Software product lines and variability modeling : A tertiary study

Raatikainen, Mikko

2019-03

Raatikainen , M , Tiihonen , J & Männistö , T 2019 , ' Software product lines and variability modeling : A tertiary study ' , The Journal of Systems and Software , vol. 149 , pp. 485-510 . https://doi.org/10.1016/j

http://hdl.handle.net/10138/299972
https://doi.org/10.1016/j.jss.2018.12.027

The Journal of Systems and Software 149 (2019) 485–510

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Software product lines and variability modeling: A tertiary study

Mikko Raatikainen [a,b,*], Juha Tiihonen [b], Tomi Männistö [b]

[a] Department of Computer Science, Aalto University, Finland
[b] Department of Computer Science, University of Helsinki, Finland

ARTICLE INFO

Article history:
Received 6 August 2017
Revised 17 November 2018
Accepted 26 December 2018
Available online 27 December 2018

Keywords:
Software product line
Variability
Variability modeling
Systematic literature review
Mapping study
Tertiary study

ABSTRACT

Context: A software product line is a means to develop a set of products in which variability is a central phenomenon captured in variability models. The field of SPLs and variability have been topics of extensive research over the few past decades. Objective: This research characterizes systematic reviews (SRs) in the field, studies how SRs analyze and use evidence-based results, and identifies how variability is modeled. Method: We conducted a tertiary study as a form of systematic review. Results: 86 SRs were included. SRs have become a widely adopted methodology covering the field broadly otherwise except for variability realization. Numerous variability models exist that cover different development artifacts, but the evidence is insufficient in quantity and immature, and we argue for better evidence. SRs perform well in searching and selecting studies and presenting data. However, their analysis and use of the quality of and evidence in the primary studies often remains shallow, merely presenting of what kinds of evidence exist. Conclusions: There is a need for actionable, context-sensitive, and evaluated solutions rather than novel ones. Different kinds of SRs (SLRs and Maps) need to be better distinguished, and evidence and quality need to be better used in the resulting syntheses.

1. Introduction

Software product line (SPL) engineering is the development of a set of products from a reusable set of assets following a common architecture and a predefined plan (Bosch, 2000; Clements and Northrop, 2001; Pohl et al., 2005). The products of an SPL can be any kinds of software systems such as embedded systems, software products or digital services. SPLs have become an important and popular means of enhancing quality, supporting reuse, and deriving different product variants efficiently. The years of research and practice have elevated variability management to become a central concern related to SPLs (Galster et al., 2014). Variability is defined as the ability of a software system or artifact to be efficiently extended, changed, customized, or configured for use in a particular context (Svahnberg et al., 2005). Variability is explicated in a software *variability model*, which we use broadly to refer to any kind of artifact abstracting or documenting variability. For example, a feature model (Kang et al., 1990; Benavides et al., 2010) is probably the first and best-known example of a dedicated variability model, but there are other similar dedicated variability models, extensions for existing models such as variability stereotypes for UML, and informal in-house developed or *ad hoc* approaches to variability modeling.

Systematic reviews (SRs) are specific classes of literature reviews that summarize and synthesize evidence about a specific topic following a predefined, systematic, and reliable research method (Dybå et al., 2005). Two typical forms of SRs are a *systematic literature review* (Kitchenham, 2007) and a *systematic mapping study* (Petersen et al., 2008), for which we use the abbreviations *SLR* and *Map*, respectively. These SRs are *secondary studies* that review original research results as their primary studies. SLRs "identify, analyse and interpret all available evidence related to a specific research question in a way that is unbiased and (to a degree) repeatable" (Kitchenham, 2007). SLRs aim to establish the state of evidence on a topic (Petersen et al., 2008). This is important for evidence-based software engineering, the aim of which is to improve decision-making related to software development and maintenance by integrating the current best evidence from the research with practical experience and human values (Dybå et al., 2005). A major step in this process is to critically appraise the evidence for its validity, impact, and applicability (Dybå et al., 2005). A Map is a "method to build a classification scheme and structure a software engineering field of interest. The analysis of results focuses on frequencies of publications for categories within the scheme" (Petersen et al., 2008). Compared to SLRs, Maps may be based on a larger number

* Corresponding author at: Department of Computer Science, Aalto University, P.O.Box 15400, FI-00076 Aalto, Finland.
E-mail addresses: mikko.raatikainen@aalto.fi (M. Raatikainen), juha.tiihonen@helsinki.fi (J. Tiihonen), tomi.mannisto@helsinki.fi (T. Männistö).

https://doi.org/10.1016/j.jss.2018.12.027

of papers and may chart research about wider topic areas, e.g., in terms of the type of contribution, such as a method or tool, and the type of research or evidence, such as a solution proposal or evaluation research (Petersen et al., 2008). However, a Map does not need to collect the empirical evidence or in-depth analysis of the subject matter due to shallow data extraction (Petersen et al., 2008). Both forms of SRs have become increasingly popular research methods (Da Silva et al., 2011) and, in practice, both forms of SRs share many similarities; and given the way they are currently applied, the distinction between the two is not always clear.

Tertiary studies are another increasingly popular form of SR that summarize the existing secondary studies. That is, a tertiary study is a review that treats other secondary studies as its primary studies and is otherwise carried out as an SLR (Kitchenham, 2007). Tertiary studies have been conducted in software engineering about three topics. First, there are studies about SRs in general. One tertiary study searched all SRs in software engineering (Kitchenham et al., 2009) and was later extended (Da Silva et al., 2011; Kitchenham et al., 2010). Second, there are studies of the SR method itself (Zhang and Ali Babar, 2013), including synthesis methods (Cruzes and Dybå, 2011), experiences (Hanssen et al., 2011; Kitchenham and Brereton, 2013), motivation (Santos and da Silva, 2013), the nature of research questions (da Silva et al., 2010), the use of SRs (Santos et al., 2014), and quality assessment in SRs (Zhou et al., 2015). Third, there are studies about specific topics, such as global software engineering (Hanssen et al., 2011; dos Santos et al., 2012; Marques et al., 2012), risks (Verner et al., 2014), requirements engineering (Bano et al., 2014), and usable results for teaching (Budgen et al., 2012). In addition, there are studies reporting the experiences of conducting SRs and discussing the SR methodology (Kitchenham and Brereton, 2013).

The fields of SPLs and variability have been subject to such extensive research that we considered it important to characterize the state of the art and practice more broadly than what is possible with an SLR or a Map. In addition, we aimed at accumulating empirical evidence from the research results for the field. Therefore, we carried out a tertiary study of SLRs and Maps. Although SPLs and variability have been the topics of several SRs, a specific tertiary study characterizing these SRs was still missing. However, there is a bibliometric analysis (Heradio et al., 2016) that focuses not only on publications but also on research topics in publications over the years. Very recently, a tertiary study (Marimuthu and Chandrasekaran, 2017) similar to ours was published, that roughly covers only the first of our three research questions. The results of both of these studies are similar to ours for the parts that overlap. Specifically, in this paper, we present the publication and general characteristics of existing SRs about SPLs and variability, an examination of how the quality of and evidence in the primary studies are assessed in these SRs, and an overview of variability models as well as supporting evidence for them in SRs. We followed the research methodological guidelines for a tertiary study (Kitchenham, 2007).

This paper is structured as follows. Section 2 introduces and motivates the research questions and the research method. Section 3 represents the full data that we extracted to the research questions. Section 4 summarizes the answers to the research question. Section 5 discusses the limitations and validity of the study. Section 6 presents the conclusions.

2. Research questions and method

2.1. Research questions

RQ1. What are the characteristics of published SRs?

- RQ1.1. What are the overall publication characteristics of SRs?
- RQ1.2. What kinds of SRs are published?
- RQ1.3. What research topics are addressed?
- RQ1.4. What is the research methodological quality of SRs that have been conducted?

RQ2. How do SRs address the quality of and evidence in the primary studies?

- RQ2.1. How is the quality of primary studies assessed in SRs?
- RQ2.2. How is the empirical evidence provided in primary studies assessed in SRs?
- RQ2.3. What kinds of evidence-based findings do SRs present?

RQ3 How is software variability modeled?

- RQ3.1. What kinds of software variability models exist?
- RQ3.2. What evidence is provided for the software variability models?

Our first research question concerns the characteristics of SRs that have been conducted about SPLs and variability in order to get an overall understanding of the conducted SR research—similarly to a Map. We characterize the publications of SRs, the SRs themselves in terms of their types and the number of primary studies they analyze, the research methodological quality of the SRs, and the topics of the SRs.

The second research question examines how SRs are being applied as a research methodological tool by focusing on the analysis and use of quality of and evidence in the primary studies of the SRs. The motivation for this is that although many concerns, especially in the early stages of the research process such as search, are addressed in the SR methodology (Kitchenham and Brereton, 2013), the analysis and use of the quality of and evidence in primary studies are not covered extensively. In fact, for evidence-based software engineering, it is essential for SRs to embrace evidence. This analysis also provides an essential basis for the third research question.

The third research question focuses on variability models, which we selected as the specific sub-topic for our study. The reason we decided to apply a tertiary study to variability models, rather than carrying out a specific SR, was to get a broader overview of variability models throughout subareas and life-cycle phases than what would be feasible in a dedicated SR about variability models. The first sub-question is aimed at getting an overview of existing variability models, while the second sub-question applies the results of the second research question to assess the evidence about variability models.

2.2. Study selection criteria

We searched for articles that fulfilled all the inclusion criteria and not any of the exclusion criteria. We included articles for which an SR was only one element of the article. The inclusion criteria were:

1. Articles that reported an SR.
2. Articles that mentioned SPL or variability as their topic using these terms or their synonyms.
3. Articles that were published in, or accepted to, a workshop, conference or journal, or were peer-reviewed technical reports or book chapters.
4. Articles that were written in English.

The exclusion criteria were:

- Master's and doctoral theses.
- Journals appearing in Beall's list and without better evidence of their scientific standard or quality.[1]

[1] The list was used as an indicator for quality of publications, and publications appearing in the list were warned to be suspicious by publication forum listing of the Finnish scientific community. However, the listing has now been discontinued.

- Informal literature reviews (lacking systematicity or at least not reporting the activities carried out, resulting in a lack of replicability in the research process).
- Articles that only discussed performing SRs but did not report the results of an original SR.
- Articles that were only related to teaching or education but did not report the results of an SR that had been conducted.
- Editorials, introductions, summaries of workshops, etc., that did not report an original SR.

2.3. Search process

The summary of the search process, which was conducted in three major phases, is given in Fig. 1. Searches of phases 1 and 2 were carried out in June and July 2015, and phase 3 was conducted in September and October 2018. For the identification of the papers, we applied four search strategies: Backward and forward searches of references — known as *snowballing search* — based on Wohlin (2014) criteria, database searches, and manual searches.

2.3.1. Snowballing searches

The snowballing search followed an iterative protocol. For each iteration, a starting set was formulated, each paper in the starting set was backward and forward snowballed, and the newly included papers were added to a starting set for the next iteration. In the backward snowballing, the reference list of a paper was reviewed. Whenever a potential paper for inclusion was found, the placements of the reference in the text of the inspected paper were investigated. The investigation of the reference's locations could provide further justification for inclusion or help to identify other similar papers. The forward snowballing was carried out using the Scopus database to find papers that cited the investigated paper. Sometimes the Scopus database did not contain the investigated paper at all or properly. In such cases, Google Scholar was used. In both cases, whenever a new candidate paper was found, the abstract of the paper was analyzed. If the paper was still considered relevant, the full paper was downloaded and evaluated for potential to be included. The first part of the snowballing search was itself a snowballing search for tertiary studies to identify the starting set of SRs for snowballing (Phase 1a in Fig. 1): A starting set of 11 tertiary studies was first identified quite unsystematically by combining the ones already known to us and by searching for additional ones using Google Scholar. The tertiary studies were forward snowballed and backward snowballed so that only other tertiary studies were considered. A total of 19 tertiary studies were identified until no new tertiary study was found after two iterations.

In the references of the 19 tertiary studies, 23 unique potential SRs were identified for inclusion. They formed the starting set of SRs, which was forward and backward snowballed until no new SR was found (Phase 1b in Fig. 1). All these preliminarily included papers were snowballed in both directions until no new SRs were found after four iterations. Again, only SRs were considered in the snowballing, not all references.

The snowballing searches identified 42 highly potential SRs for inclusion. One paper that was published in a journal that appeared on Beall's list was excluded. A citation matrix about backward and forward snowballing was created to verify the completeness of snowballing in both directions.

In order to validate the search, we applied a quasi-gold standard, meaning that we compared our search results with a set of SRs about the research topic (Zhang et al., 2011). The set of 26 SRs in our earlier records was used as a quasi-gold standard rather than the recommended method based on a manual search. The snowballing search had revealed 25 of these 26 SRs. Thus, the quasi-sensitivity was 96%, exceeding the "high sensitivity" class (85–90%) (Zhang et al., 2011). The SR that was not found was added and snowballed, and thus a set of 43 SRs were potentially included, i.e., they seemed to fulfill the inclusion and exclusion criteria. Nevertheless, we decided to pilot additional searches because our quasi-gold standard was not formed as rigorously as recommended, and out of general interest. Because we found several new SRs in the pilot, we decided to design and carry out database and manual searches.

2.3.2. Database and manual searches

The database searches (Phases 2 and 3 of Fig. 1) were carried out in five databases: Web of Science (WoS), Scopus, IEEEXplorer, ACM Digital Library, and Science Direct. The search that we adapted to each database consisted of five parts, as described below and exemplified for Scopus using the same parts in Fig. 2:

1. The search targeted metadata — or abstract, title, and keywords — depending on what was possible in the specific database.
2. The terms used to find SRs combined all general search terms found from the eleven tertiary studies known at the time that explicated their search terms.
3. The search was limited to the field of software.
4. SPL and variability were used as search terms, and we considered synonyms for these terms.
5. The search was limited to computer science or similar subject areas, if possible.

For the search results of each database search, we inspected the title and publication. If the paper seemed to be potentially relevant to SPLs or variability, we read the abstract. We did not exclude papers only on the basis of not mentioning a SR in the title. We excluded a paper if the combination of the title and publication was deemed to be out of scope. For example, the search results included several papers whose titles were too generic to make a decision, but these papers were published in journals completely unrelated to software, such as agricultural journals, which resulted in their exclusion. A manual search process addressed the forums devoted to the SPLs: The Software Product Line Conference (SPLC) and SPLC workshops, the International Workshop on Variability Modelling of Software-intensive Systems (VaMoS), and the International Workshop on Variability in Software Architecture (VarSa). Thus, the manual search did not cover journals. The basic bibliographic information of all papers in these venues was collected on a spreadsheet. A paper was considered for inclusion if the title suggested a potential SR. However, all potential papers identified in the manual search were already included or excluded. Finally, as a central forum for publishing SRs, we read manually the titles of International Conference on Evaluation and Assessment in Software Engineering (EASE) papers from 2007 to 2015, which are accessible online, but no new candidates were found.

Finally, backward and forward snowballing was carried out for the newly found SRs from the database and manual searches similarly to the process described above but no new SRs were found. However, we did not assess the database and manual search results for validity.

The 43 papers from the snowballing search and the papers from the database search were combined. After the removal of duplicates and the application of the inclusion and exclusion criteria, we had identified 59 SRs for inclusion.

2.3.3. Phase 3: Update to include the most recent papers

The third phase of searches (Phase 3 of Fig. 1) was conducted with a similar protocol than earlier to cover the most recent publications. First, a database search was conducted. The same database search protocol described above was followed, with an additional limitation to publication years 2015–2018, and after applying the inclusion and exclusion criteria 26 new SRs were included. These

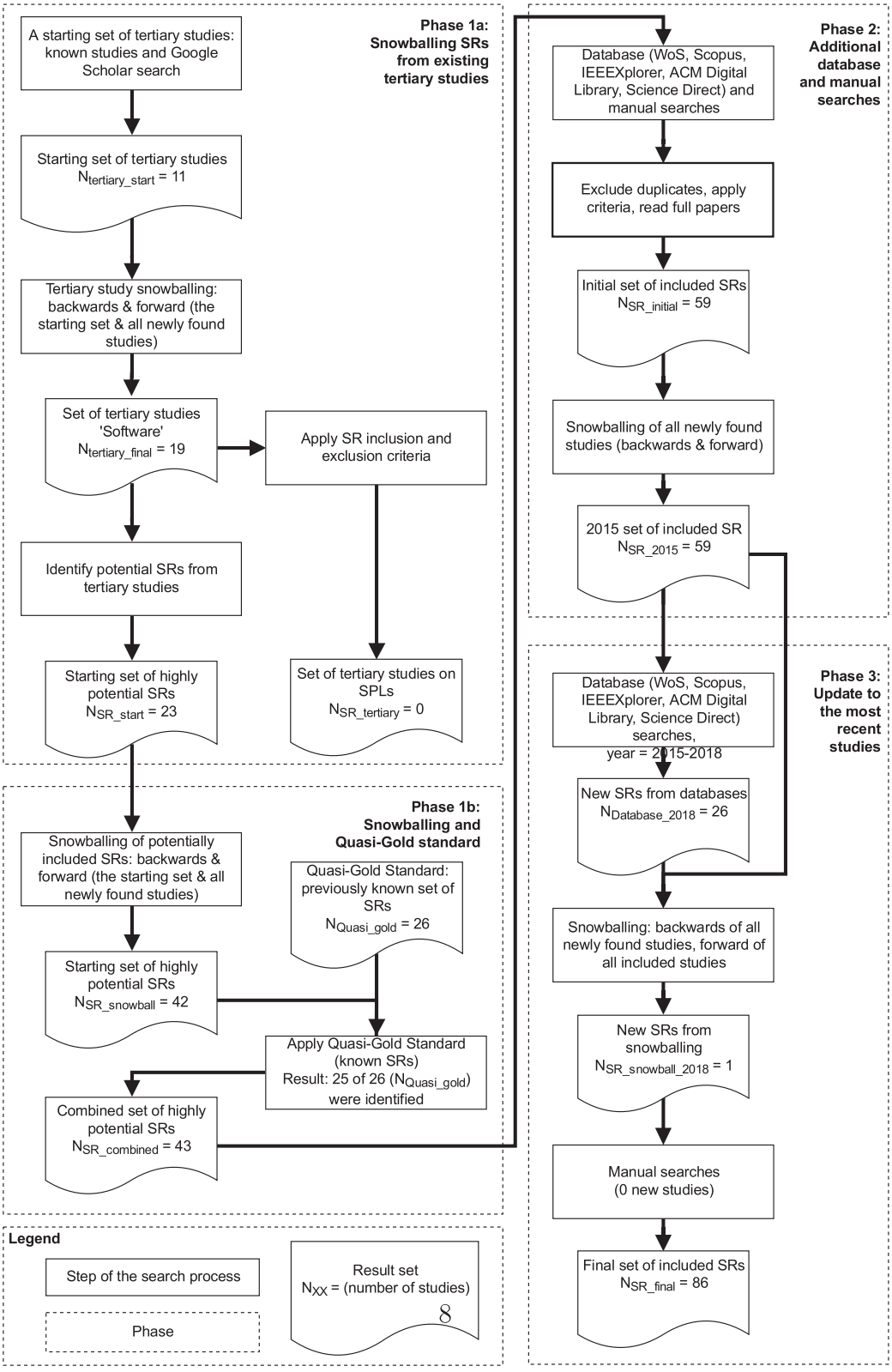

Fig. 1. The activities and results of the search process, which consisted of three phases.

new SRs were snowballed using the same protocol as earlier. In addition, the set of 59 earlier SRs were forward snowballed for citing papers from published in 2015 or later. As a result of snowballing, one new SR was included, which was likewise snowballed. The snowballing resulted in only one iteration. Manual searches targeted the above mentioned publication venues in 2015–2018; no new SRs were included. As a result, the final set of included SRs has 86 studies.

The number of SRs found in different stages is given in Table 1. The 'Total' column is the number of potential SRs found from the

```
1: TITLE-ABS-KEY (
2:     ("analysis of research" OR "body of published knowledge" OR "body of published research
       " OR "controlled review" OR "controlled survey" OR "empirical body of knowledge" OR
       "evidence based" OR "evidence-based" OR "exhaustive review" OR "exhaustive survey" OR
       "in-depth survey " OR "literature analysis" OR "literature review" OR "literature search" OR
       "literature survey" OR "mapping" OR "mapping studies" OR "mapping study" OR "meta
       analysis" OR "meta-analysis" OR "overview of existing research" OR "past studies" OR "re-
       view of studies" OR "scoping study" OR "SLR" OR "structured review" OR "structured
       survey" OR "study aggregation" OR "study synthesis" OR "subject matter expert" OR "sys-
       tematic literature review" OR "systematic map" OR "systematic mapping" OR "systematic
       review" OR "systematic survey")
3:     AND ("software")
4:     AND ("product line" OR "product family" OR "product lines" OR "product families" OR
       "variability" OR "variabilities"))
5:     AND (LIMIT-TO (SUBJAREA , "COMP"))
```

Fig. 2. An example of the database search string applied to the Scopus database.

Table 1
The number of SRs found in different phases and stages from the sources.

Stage	Snowballing[a]	Scopus	ACM	WoS	IEEE	Science Direct
Initial search (Phases 1 & 2)						
Total	47	348	140	218	105	15
Full text analysis	42	50	41	35	16	6
similar Included	42	39	30	26	10	6
Unique	7	5	3	1	2	0
Search update(Phase 3)						
Total	1[b]	145	52	70	40	55
Full text analysis	1	33	10	20	6	11
Included	1	9	7	3	5	2
Unique	1	2	2	0	0	1

[a] These figures do not include the SR that we added after applying the quasi-gold standard. The added SR was also found in Scopus and WoS.
[b] We did not include SRs that were already found in earlier stages.

Table 2
Data extraction form.

Data	RQ
Full bibliographic information	1.1
Publication type *(Journal/conference/workshop/other)*.	1.1
Type of SR *(Map/SLR)*	1.2
Number of primary studies	1.2
Topic	1.3
Is there an explicit difference between domain engineering and/or application engineering in separate sections or the research questions; or implicit or generic in this sense *(specific/generic)*?	1.3
Was the domain engineering phase covered explicitly, partially so that there was no explicit claim of the phase but the relevant phase can be deduced, or not at all or not in an identifiable manner *(yes/partially/no)*?	1.3
Was the application engineering phase covered explicitly, partially so that there was no explicit claim of the phase but the relevant phase can be deduced, or not at all or not in an identifiable manner *(yes/partially/no)*?	1.3
Did the paper address tools either as the topic of paper, in the research questions, or as a topic in the analysis *(Topic of the paper/A research question/Analysis topic)*?	1.3
QA scores for QA1-7 *(yes/partially/no)*, see Table 3.	1.4
Reasons for not scoring 1 in QA1-7	1.4
How were the search results assessed for completeness?	1.4
What quality and evidence assessment frameworks were used for the analysis of the primary studies?	2.1&2.2
How were the quality and evidence in the primary studies used in the findings of the SR?	2.3
What variability models exist?	3.1
What evidence exists about the variability models?	3.2

source. The 'Full text analysis' column refers to the number of SRs that were downloaded and analyzed. The 'Included' column indicates the number of SRs that were ultimately included from each source. The 'Unique' column includes the number of SRs that was found only from the specific source. During the snowballing of the search update, we kept tract only new SRs.

2.4. Data extraction

The analyzed data from each SR is shown in Table 2. If only enumerated values were extracted, the possible values are given in parentheses. The 'RQ' column shows the respective primary research questions.

The extracted full bibliographic information provides data about publications for RQ1.1. For RQ1.2) we identified the number of primary studies and the type of an SR. The type is either a Map or an SLR as defined by the SR, i.e., by the self-claim of the authors rather than the analysis of the content. The main topics of SRs were extracted from the title and content for RQ1.3. We also differentiated how each SR addressed domain and application engineering phases including whether the topic was specific for differ-

Table 3
The quality assessment criteria.

QA	Criteria
QA1	*Are the review's inclusion and exclusion criteria described and appropriate?* (1): The criteria are explicitly defined in the study. Our interpretation: Criteria are in a table, bullet points, or described clearly in the text. (0.5): The criteria are implicit. Our interpretation: Inclusion and exclusion was made on the basis of the research questions and search terms without being more explicit. (0): The criteria are not defined and cannot be readily inferred.
QA2	*Is the literature search likely to have covered all relevant studies?* (1): The authors have either searched four or more digital libraries and included additional search strategies or identified and referenced all journals addressing the topic of interest. Our modification: We accept also the application of an extensive backward and forward snowballing search strategy (Wohlin, 2014). (0.5): The authors have searched four digital libraries with no extra search strategies or three digital libraries with extra search strategies. (0): The authors have searched up to 2 digital libraries or an extremely restricted set of journals.
QA3	*Did the reviewers assess the quality/validity of the included studies?* (1): The authors have explicitly defined quality criteria and extracted them from each primary study. Our interpretation: The quality of the primary study is assessed, such as the appropriateness of the study design. (0.5): The research question involves quality issues that are addressed by the study. Our interpretation: Only the study design of primary studies is addressed but not assessed for quality/validity. For example, the applied research method is extracted, or the evidence is characterized. (0): No explicit quality assessment of individual primary studies has been attempted.
QA4	*Were the basic data/studies adequately described?* (1): Information is presented about each study. Our interpretation: References to the primary studies are provided. (0.5): Only summary information about primary studies is presented. Our interpretation: The number of primary studies was mentioned. (0): The results of the individual primary studies are not specified. Our interpretation: Not even the number of primary studies supporting claims was explicated.
QA5	*Is the analyzed quality of primary studies traceable to primary studies?* (1): Fully traceable to primary studies. (0.5): Only part of the quality is traceable, quality is traceable only for some of the primary studies, or only numeric summaries are given. (0): Not traceable.
QA6	*Is the identified evidence traceable to primary studies?* (1): Fully traceable to primary studies. (0.5): Only part of the evidence is traceable, evidence is traceable only for some of the primary studies, or only numeric summaries are given. (0): Not traceable.
QA7	*Is the analyzed quality and evidence used in the findings?* (1): Findings are reported systematically so that at least the best quality or evidence is taken into account. (0.5): Quality or evidence is not explicitly used in the findings but somehow related to the findings: Studies with the best quality or evidence are identified but not described, or quality or evidence is used in a non-traceable manner, such as numerical summaries. (0): Neither quality nor evidence is used.

ent phases or generic. The dichotomy of domain and application engineering is typical in SPL engineering and is roughly paralleled to development for reuse or core asset development, and development with reuse or product development, respectively. As an additional concern, the role of tool support was extracted.

To assess the research methodological quality of the conducted SRs (RQ1.4), we first applied the four York University DARE criteria for quality assessment (QA1-4), following the practices in other tertiary studies (Kitchenham et al., 2009; 2010; Da Silva et al., 2011; Bano et al., 2014; Verner et al., 2014). Some tertiary studies, including ours, have modified and interpreted the wording of the original criteria in (Kitchenham et al., 2009). Our interpretations and adaptations are explained within the criteria of the quality scores in Table 3. Additionally, Table 3 includes three new quality criteria (QA5-7) that we developed to characterize the analyses of quality of and evidence in the primary studies, and the use of the results of these analyses in the findings. Although Maps may only provide an overview of the topics addressed in the research, we analyzed all QAs in all SRs. The quality scores applied for every QAs were:

1. 1 if the criterion was fulfilled fully,
2. 0.5 if the criterion was fulfilled partly, and
3. 0 if the criterion was not met.

In addition to determining the score for the QA criteria, we carried out an additional analysis about the QA criteria related to RQ1.4. We recorded the reasons why an SR did not reach a score of 1 for all QAs. We also studied how the search results were assessed for completeness or validity. The addition was made because we did not want to modify QA2 significantly or to add another QA for search criteria.

For RQ2.1 and RQ2.2, we collected the frameworks used in the SRs to analyze the quality of or evidence in the primary studies. For RQ2.3 we analyzed how the SRs use the analyzed quality and evidence.

The variability models for RQ3.1 were extracted from the SRs so that original terminology was primarily used. To provide data for RQ3.2, we analyzed the quality and evidence of the supporting primary studies (cf. RQ2.1 and RQ2.2) for each variability model.

The first author extracted all data, and the second author checked the data extractions of a few SRs that he selected. In addition, whenever the first author had uncertainties about the analysis, he discussed it with the second author until an agreement was reached. Finally, the second author checked the QA scores of the SRs that received the best scores for QA3 or QA5-7.

2.5. Data analysis and presentation

We applied spreadsheets extensively to store the extracted data presented in Table 2. Section 3 presents the full data and corresponding direct results to each research question. Section 4 summarizes the answers to the research questions and presents characterizations, trends, and individual findings on the basis of further analyses. We also present the cross-analysis of the data over different research questions.

3. Results: Data on the research questions

This section provides basic data on the research questions. The full list of included SRs is given in Appendix. The SRs are referred as S1-S86. In the case of a few SRs, at least partially the same authors carried out an extension of their previous work. S13 extended the analysis for the same set of primary studies as S11, and

Table 4
Extracted basic data of SRs.

ID	Year	Pub[a]	Type	S/G[b]	DE[c]	AE[d]	Tool[e]	#[f]	Topic
S1	2010	J	SLR	g	y	n	RQ	49	Requirements engineering
S2	2014	W	Map	g	p	n	–	62	Feature location for SPL adoption
S3	2015	J	SLR	g	Y	n	RQ	13	Feature extraction from requirements
S4	2014	C	Map	g	y	p	RQ	17	Measures in feature model
S5	2010	W	Map	g	p	y	A	29	Dynamic SPL
S6	2007	C	SLR	g	y	p	–	8	Design patterns
S7	2012	J	SLR	s	y	y	–	41	Testing strategies
S8	2014	J	SLR	s	y	y	A	49	Testing strategies
S9	2014	J	Map	g	y	y	RQ	58	Service oriented SPL
S10	2009	C	SLR	g	p	p	A	34	Variability management
S11	2009	C	SLR	g	y	p	–	97	Research methods in variability management
S12	2009	W	SLR	s	y	n	–	19	Scalability in variability modeling
S13	2011	J	SLR	g	y	p	–	97	Research methods in variability management
S14	2011	J	SLR	s	y	y	–	39	Agile methods
S15	2011	J	Map	g	y	y	A	64	Testing
S16	2011	C	Map	g	n	n	–	34	Adoption
S17	2014	C	SLR	s	y	y	RQ	19	Embedded systems using MDE and SPL
S18	2014	J	SLR	s	y	p	A	196	Variability management
S19	2015	C	SLR	g	y	y	–	13	Variability in architecture knowledge management
S20	2013	J	SLR	g	p	n	–	45	Feature model analysis
S21	2012	J	SLR	s	y	y	–	37	Multi SPL
S22	2010	W	SLR	g	y	y	–	16	Research methods in feature modeling
S23	2013	J	SLR	g	p	p	RQ	127	Separation of concerns in feature modeling
S24	2011	W	SLR	g	y	y	–	3	Research methods in testing
S25	2008	W	SLR	g	n	n	–	19	Research methods in SPL economics
S26	2009	J	SLR	g	n	n	–	89	Research methods in domain analysis
S27	2014	J	SLR	s	y	y	RQ	20	Traceability
S28	2013	J	Map	g	y	n	RQ	74	Adoption: Reengineering and refactoring
S29	2009	C	SLR	g	y	y	A	23	Testing
S30	2013	C	SLR	g	y	y	A	37	Testing
S31	2012	C	SLR	s	y	y	A	25	Testing
S32	2010	J	SLR	s	y	y	T	19[g]	Domain analysis tools
S33	2013	J	Map	g	y	p	–	30	Risk management
S34	2015	J	Map	s	y	y	A	77	Search based software engineering for SPL
S35	2015	W	Map	g	y	n	–	47	Combinatorial interaction testing
S36	2013	J	SLR	g	y	n	A	46	Variability in quality attributes in service SPL
S37	2013	J	SLR	g	p	p	A	81	Service oriented SPL
S38	2009	C	SLR	s	y	y	–	39	Quality attributes
S39	2012	J	SLR	s	y	y	A	35	Quality attributes and measures
S40	2009	W	SLR	g	p	n	–	13	Scoping
S41	2011	C	SLR	g	n	n	A	43	Quality attributes
S42	2009	C	SLR	g	p	p	–	7	Mobile middleware SPL
S43	2011	J	Map	g	p	p	–	45	Testing
S44	2011	W	Map	g	y	y	A	48	Service oriented SPL
S45	2012	C	SLR	g	y	n	–	29	Quality attribute variability
S46	2012	C	Map	g	n	n	T	9	Testing tools
S47	2012	C	SLR	g	p	n	–	57	Evaluation
S48	2014	C	SLR	s	y	y	T	41[g]	Management tools
S49	2010	J	SLR	s	y	p	A	118	Derivation support
S50	2014	C	Map	s	y	n	–	9	Variability in textual use cases
S51	2015	C	Map	g	p	n	A	24	Consistency checking
S52	2013	J	SLR	s	y	y	–	63	Service oriented SPL
S53	2011	J	Map	g	p	p	–	32	Agile methods
S54	2013	C	SLR	g	y	y	A	20	Dynamic SPL derivation
S55	2014	C	SLR	g	p	p	–	36	Quality attributes
S56	2008	C	SLR	s	y	n	–	17	Domain design
S57	2015	J	SLR	g	n	n	–	31	Adoption
S58	2012	W	SLR	g	p	n	–	30	Service identification in SPL
S59	2014	C	SLR	g	p	p	–	18	Bad Smells
S60	2016	J	Map	g	p	p	A	19	Intelligent configuration
S61	2017	J	Map	g	y	n	RQ	119	Reengineering legacy applications
S62	2017	J	SLR	g	p	p	T	37	Tools for variability
S63	2016	C	Map	s	y	y	–	39	Functional safety in SPL
S64	2018	J	Map	s	y	y	RQ	47	Integration of feature models
S65	2016	C	Map	g	p	n	A	165	Change impact
S66	2016	C	Map	g	p	p	–	9	Dependability for DSPLs
S67	2018	J	SLR	g	p	p	–	42	Variability metrics
S68	2018	J	Map	g	y	y	–	423	Feature model analysis
S69	2015	C	Map	s	y	y	A	54	Variability management in DSPLs
S70	2017	C	Map	g	n	n	–	68	Automotive agile SPLs
S71	2018	C	Map	g	n	n	–	66	Agile transformation and SPL
S72	2017	C	SLR	g	p	n	RQ	25	Reverse engineering variability from natural language
S73	2015	C	SLR	g	p	n	A	28	SPL architecture recovery
S74	2016	C	SLR	g	p	n	A	35	SPL architecture metamodels
S75	2016	C	Map	s	y	y	RQ	32	Visualization for SPLs
S76	2018	J	Map	s	y	y	RQ	37	Visualization for SPLs

(*continued on next page*)

Table 4 (*continued*)

ID	Year	Pub[a]	Type	S/G[b]	DE[c]	AE[d]	Tool[e]	#[f]	Topic
S77	2018	J	SLR	g	p	n	RQ	60	SPL evolution
S78	2016	J	MAP	s	y	y	–	107	SPL evolution
S79	2018	J	SLR	s	n	y	A	66	Configuration of extended feature models
S80	2016	J	Map	g	p	p	–	44	Combinatory integration testing
S81	2016	J	SLR	g	p	n	A	54	Requirements models in SPLs
S82	2016	C	SLR	s	y	y	–	37	Requirements engineering and variability in DSPLs
S83	2018	C	Map	g	p	n	–	35	Architecture recovery for SPLs
S84	2018	J	Map	s	y	y	A	35	Feature interaction
S85	2015	J	SLR	g	n	n	–	15	Effects on experience
S86	2017	J	Map	s	y	y	–	62	Traceability

[a] Type of publication: Journal (J), conference (C), workshop (W).
[b] Differentiation of domain and/or application engineering: Specific (s), general not explicit (g).
[c] Applicable to domain engineering: Yes (y), partially (p), no (n).
[d] Applicable to application engineering: Yes (y), partially (p), no (n).
[e] The role of tools: Topic of the paper (T), a research question (RQ), analysis topic (A).
[f] The number of primary studies.
[g] The study reported the number of tools, not the number of primary studies.

S30 and S8 extend the previous searches to cover recent years for the same research questions. S76 extends S75, and S61 extends S2 by refined search and additional research questions. We interpret that S37 extends S44, e.g., by refined research questions, searches, and analyses, although few details about the possible extension are given. However, we did not exclude these extensions from the final set of SRs.

3.1. Data on RQ1.1 and RQ1.2: SRs' publications and characteristics

The SRs including the full bibliographic details are listed in the Appendix. The extracted basic data is shown in Table 4.

3.2. Data on RQ1.3: Addressed topics

The topics of SRs are presented in the 'Topic' column of Table 4 roughly as stated by each SR. Additionally, Table 4 presents how the domain and application engineering, and tools are covered by each SR.

In order to group similar topics and provide a better overview, we mapped the SRs to more general *topic categories* as shown in Table 5. One SR can appear in more than one category. The topic categories were formed in a bottom-up manner by grouping similar SRs and assigning each to one or more categories. The categories follow the commonly applied software engineering categorization that is applied, e.g., in SWEBOK (Abran et al., 2004). The first three topic categories distinguish between the common software engineering life cycle phases of *requirement engineering*, architecture and detailed *design*, and *testing*. The table omits *implementation*, because it was not addressed by any SR. Crosscutting topics over life-cycle phases, i.e., *variability management, quality attributes*, and *software process model*, were distinguished as categories of their own. Practically all papers deal with variability in some way. Therefore, to classify an SR into the variability management topic category, we required a specific variability management topic such as addressing variability or feature modeling. *Management* goes over life-cycle phases including SPL adoption, economic concerns, risk management, and evolution. *Specific SPLs* focus on a specific application domain or a specific kind of SPL. These are the service-based systems (7 SRs), dynamic SPLs (5), multi SPLs, embedded SPLs, automotive agile SPLs, and mobile middleware SPLs. Finally, *research methods* means that the SR has a significant focus on assessing, e.g., which research methods the primary studies applied or how well the methods were applied.

3.3. Data on RQ1.4: Quality assessment of SRs

Table 6 conveys the full quality assessment scores. The reasons for the SRs not scoring 1 in each QA are elaborated below.

QA1: The inclusion and exclusion criteria. One SR does not provide enough details to interpret the inclusion and exclusion criteria, thus scoring 0. Twenty SRs score 0.5 because the inclusion and exclusion criteria are not reported explicitly but the studies report enough details about the search process, such as research questions and search strings, so that it became almost clear how studies were found.

QA2: The completeness of the literature search. Nine SRs score 0: S9 used only Google Scholar search to a selected set of publishers; S23 used one database and its own records; S49 carried out automated searches only for a set of selected publications; S56 used web search and manual search for selected publications; S5, S22, and S45 targeted the search only to a limited set of conferences; and S68 and S69 searched only two databases and used limited snowballing. Twenty-six SRs score 0.5, out of which 18 relied only on the database searches and lack manual searches, and seven SRs (S6, S15, S18, S21, S38, S40, S48) searched a limited set of databases. S6 has both of these deficiencies. S31 relied largely on the previously performed searches that have these deficiencies.

Search completeness beyond QA2: Ten SRs assessed search completeness. S36 established a quasi-gold standard by manual search, and S15 used another SR as a quasi-gold standard. S10, S11, and S13 checked if three relevant known papers were found, S63 and S68 checked four known papers, and S81 checked five. S78 checked the 34 primary studies of another SR and found 9 new papers, but did not revise its search. S39 had a set of studies that were compared to the search results, but further details are not given.

A few SRs relied at least partially on previous search results but did not otherwise assess search completeness. S7 included the studies of three other SRs and added a manual and database search. S31 analyzed for the inclusion papers from existing SRs. S58 included the primary studies of an existing SLR and extended the set of studies with a manual search and a snowballing search.

QA3: The quality assessment of the primary studies. QA3 is completely omitted in 21 SRs. Thirty-eight SRs carried out a partial quality assessment of the primary studies (QA3 score = 0.5). These SRs explicitly assessed the evidence in the primary studies, such as the applied research methods or whether industrial evidence exists. The remaining 15 SRs score 1, out of which all of them analyzed both quality and evidence except for S21, S33, and S49, which analyzed quality only and applied it as an inclusion criterion.

Table 5
Topic categories of SRs over years. One SR may have several topic categories.

Topic category	#	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Requirements engineering	7			S26 S40	S1 S32				S50	S3	S81		
Design	6	S6	S56				S47		S27 S59	S19			
Testing	13			S29		S15 S24 S43	S7 S31 S46	S30	S8	S35 S51	S79 S80		
Variability management	25			S10 S11 S12	S22 S49	S13		S20 S23	S4 S18 S48	S34 S69	S60 S74 S75 S82	S62 S86	S64 S67 S68 S76 S79 S84
Quality attributes	8			S38		S41	S39 S45	S36	S55		S63 S66		
Process model	4					S14 S53						S70	S71
Management	17					S16		S28 S33	S2 S27 S48	S3 S51 S57 S73 S85	S65 S78	S61 S72	S77 S83
Specific SPLs	15			S42	S5	S44	S21 S58	S36 S37 S52 S54	S9 S17	S69	S66 S82	S70	
Research methods	4		S25	S11 S26		S13							

QA4: The basic data of the primary studies. One SR scores 0, meaning that only a summary is given without even the number of supporting primary studies. Out of the five papers that score 0.5, three (S11, S38, S44) provide only the number of primary studies supporting the findings without referencing the primary studies. In S5, it remained unclear what papers are included. S9 lists but does not make references in the text to the primary studies.

QA5: The quality traceability of the primary studies. Sixty-five SRs score 0 because they do not report the quality of individual studies. Seven SRs score 0.5: S8 provides a graph of scores without traceability to primary studies. S86 uses a bubble chart to visualize the number of papers with a specific combination of rigor and relevance scores. S18 and S62 represent only a numeric summary. S11 and S13 analyzed only those primary studies that they considered empirical, providing for each study a summary and full score, respectively. S55 reports a numeric summary for four out of the eight applied criteria. Four out of the remaining eight SRs scoring 1 (S7, S14, S21, S49) used quality assessment only as an inclusion criterion meaning that their quality assessment criteria are fulfilled although actual scores are not given.

QA6: The evidence traceability of the primary studies. Evidence is absent in 24 SRs (score 0). Partial traceability is provided by 23 SRs (score 0.5). S35 lists only a subset of evaluation, and S54 does not state unambiguously the evidence. The remaining SRs scoring 0.5 provide a numeric summary of primary studies in different evidence categories without references to primary studies and, thus, without traceability.

QA7: The use of evidence or quality in findings. QA7 score is zero in 47 SRs. Thirty-two SRs score 0.5: Several SRs (S8, S15, S19, S29, S30, S32, S39, S56, S59, S67, S69, S73, S74, S80, S81) provide listings from which the level or nature of evidence in the primary studies can be observed for different topics. Many SRs (S4, S5, S9, S16, S28, S34, S37, S43, S50, S51, S53, S65, S79, S86) present the results in a two-dimensional plot such as a bubble plot: The topics or research questions are on one axis, the level of evidence on the other axis, and the number of studies are then plotted in this space. S15, S28 and S51 organize the evidence by different topics. S8 differentiates a numeric summary by topic. S55 and S86 identify the primary studies with the best evidence. S18 and S36 identify the primary studies of best rigor and relevance by combining the quality, evidence, topics of these studies.

3.4. Data on RQ2.1: Quality assessment of primary studies

The *quality questions* (Dybå and Dingsøyr, 2008), which have been developed for assessing the primary studies in software engineering SRs, are applied in 13 SRs (the 'Quality Questions' row of Table 7). The original 11 questions (Table 8) address four concerns: Reporting (questions 1–3), rigor (4–8), credibility (9–10), and relevance (11). All questions are not always used except in S13 and S52 and sometimes there are minor wording changes. In addition, five SRs add extensions as shown in Table 8. The reasons for omitting and extending the questions are not elaborated. S8 used the first quality question to exclude non-empirical primary studies.

Ten SRs apply other means for the quality assessment than the quality questions (the 'Other quality assessment' row of Table 7). S43, S54, S62, S63, S70, S75 have their own sets of questions and S33 adds several questions to the quality questions. These questions have some similarities with the quality questions, such as a few same or very similar questions, but the other questions are still quite unique. S7, S21, and S49 specify their own quality criteria or questions that are used in the inclusion and exclusion criteria. S18, S36 and S88 analyze rigor and relevance based on existing criteria (Ivarsson and Gorschek, 2011).

Table 6
QA scores for the SRs.

ID	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA sum
S1	1	1	1	1	1	1	0	6
S2	1	.5	.5	1	0	.5	0	3.5
S3	1	1	1	1	0	1	0	5
S4	1	.5	.5	1	0	1	.5	4.5
S5	1	0	.5	.5	0	.5	.5	3
S6	.5	.5	0	1	0	0	0	2
S7	.5	1	1	1	1	0	0	4.5
S8	1	1	1	1	.5	1	.5	6
S9	.5	0	.5	.5	0	.5	.5	2.5
S10	1	1	0	1	0	0	0	3
S11	1	1	1	.5	.5	.5	0	4.5
S12	1	1	0	1	0	1	1	5
S13	1	1	1	1	.5	.5	1	6
S14	1	1	1	1	1	0	1	6
S15	.5	.5	.5	1	0	1	.5	4
S16	1	1	.5	1	0	.5	.5	4.5
S17	1	.5	.5	1	0	1	0	4
S18	1	.5	1	1	.5	.5	.5	5
S19	1	1	.5	1	0	1	.5	5
S20	1	1	0	1	0	0	0	3
S21	1	.5	1	1	1	0	0	4.5
S22	1	0	.5	1	0	1	1	4.5
S23	.5	0	0	1	0	0	0	1.5
S24	.5	1	.5	1	0	1	1	5
S25	.5	1	.5	1	0	1	0	4
S26	1	1	.5	1	0	1	0	4.5
S27	0	.5	0	1	0	0	0	1.5
S28	1	1	.5	1	0	.5	.5	4.5
S29	.5	1	.5	1	0	1	.5	4.5
S30	.5	.5	.5	1	0	1	.5	4
S31	.5	.5	.5	1	0	1	0	3.5
S32	1	1	.5	1	0	1	.5	5
S33	1	1	1	1	0	1	0	5
S34	1	1	.5	1	0	1	.5	5
S35	1	1	.5	1	0	.5	0	4
S36	1	1	1	1	1	1	.5	6.5
S37	1	1	.5	1	0	.5	.5	4.5
S38	1	.5	.5	.5	0	.5	0	3
S39	1	1	.5	1	0	1	.5	5
S40	.5	.5	0	1	0	0	0	2
S41	.5	.5	.5	0	0	.5	0	2
S42	1	.5	0	1	0	0	0	2.5
S43	1	1	1	1	0	.5	.5	5
S44	1	1	.5	.5	0	.5	0	3.5
S45	1	0	.5	1	0	1	1	4.5
S46	1	1	.5	1	0	1	0	4.5
S47	.5	.5	0	1	0	0	0	2
S48	1	.5	0	1	0	0	0	2.5
S49	1	0	1	1	1	0	0	4
S50	1	.5	.5	1	0	1	.5	4.5
S51	1	1	.5	1	0	.5	.5	4.5
S52	1	.5	1	1	1	1	0	5.5
S53	1	1	.5	1	0	.5	.5	4.5
S54	1	1	1	1	1	.5	0	5.5
S55	1	1	1	1	.5	.5	.5	5.5
S56	1	0	.5	1	0	1	.5	4
S57	1	1	0	1	0	0	0	3
S58	.5	1	.5	1	0	0	0	3
S59	1	.5	.5	1	0	1	.5	4.5
S60	.5	.5	0	1	0	0	0	2
S61	1	1	0	1	0	1	0	4
S62	1	1	1	1	.5	0	0	4.5
S63	1	1	1	1	0	.5	0	4.5
S64	1	.5	.5	1	0	1	0	4
S65	.5	1	.5	1	0	1	.5	4.5
S66	.5	.5	.5	1	0	1	1	4.5
S67	1	0	.5	1	0	1	.5	4
S68	.5	0	.5	1	0	1	0	3
S69	1	1	1	1	0	1	.5	5.5
S70	1	1	0	1	0	0	0	3
S71	1	1	0	1	0	0	0	3
S72	1	1	0	1	1	0	0	4
S73	1	1	.5	1	1	1	.5	6
S74	1	1	1	1	1	1	.5	6.5
S75	1	1	0	1	0	0	0	3
S76	1	1	.5	1	1	0	0	4.5
S77	1	.5	1	1	1	1	0	5.5
S78	1	.5	.5	1	0	1	0	4
S79	1	.5	.5	1	0	.5	.5	4
S80	1	1	.5	1	0	1	.5	5
S81	1	1	1	1	1	1	.5	6.5
S82	1	1	0	1	0	0	0	3
S83	.5	.5	.5	1	0	.5	0	3
S84	1	1	.5	1	0	.5	0	4
S85	.5	1	0	1	0	0	0	2.5
S86	1	1	1	1	.5	.5	.5	5.5
Average	0.88	0.75	0.53	0.97	0.21	0.59	0.27	4.21

(*continued on next page*)

3.5. Data on RQ2.2: Evidence assessment of primary studies

There are four different evidence assessments, each of which is applied in at least seven SRs (the second block of Table 7). The 'Other evidence assessment' row of Table 7 identifies SRs that applied an assessment that at most one other SR applies. Some SRs apply more than one assessment. The applied evidence assessments are ordered by year, which does not show any conclusive trends such as disappearance of one assessment or clear dominance of another assessment. Respectively, it is still not uncommon that no quality or evidence assessment is applied as shown in the last row of Table 7.

Research classification. The six different classes of research (Wieringa et al., 2006) were used in 20 SRs: *Solution proposals* make a proposal for something new, supported merely by a small example or a good line of argument. *Validation research* means that the object of investigation is novel but validated in academic settings rather than applied in practice. *Evaluation research* means that a study is carried out in a practical setting. *Philosophical papers* propose something significantly novel relying merely on argumentation. *Opinion papers* are about authors' opinions not based on the results of research. *Experience papers* are about the authors' personal experience such as lessons learned. The research classification is also adapted. S15 and S37 add *conceptual proposition*, S9 differentiates *early research advancement*, S44 includes *survey papers*, and S83 adds *exploratory research*.

Evidence level. The evidence level framework, which seven SRs applied, is adopted to software engineering (Kitchenham, 2004; Alves et al., 2010) and it differentiates evidence from L1 (weakest) to L6 (strongest). The levels are *L1 no evidence; L2 demonstration, toy example; L3 expert opinions, observations; L4 academic studies; L5 industrial studies*; and *L6 industrial evidence*. As most of the primary studies fall in L1 or L2, S81 refines the level 2.

Industrial or non-industrial settings. Twenty-two SRs analyzed the settings where the primary studies had been carried out. The settings include 'artificial settings', 'examples', 'scaled down real examples' 'randomized' and 'no evaluation at all'. Unfortunately, the categories between SRs are largely incommensurable. It is only possible to identify whether some real industrial settings are involved or not, which we consider as an additional classification. The presence of an industrial setting can also inferred via the other frameworks: Research classification to experience or evaluation paper, and the evidence levels L3, L5, and L6 require industrial evidence.

Research methods. Twenty SRs assessed the applied research methods in evaluating evidence. Unfortunately, the SRs' use of research method categorization is a too heterogeneous and incommensurable for us to meaningfully present and summarize them: We attempted to categorize the research methods by combining similar research methods such as 'experience' and 'expert opinion'

Table 7
The applied quality and evidence assessment over years.

	#	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Quality assessment													
Quality questions (Dybå and Dingsøyr, 2008)	13			S11	S1	S13 S14		S33 S36 S52	S8 S18 S55	S3 S81			S77
Otherquality assessment	10				S49	S43	S7 S21	S54		S69	S63 S74	S62 S86	
Evidence assessment													
Research classification(Wieringa et al., 2006)	20				S5	S15 S16 S43 S44 S53	S46	S28 S37 S54	S4 S9 S50	S51 S73	S66 S78	S86	S68 S83
Evidence level (Alves et al., 2010)	7				S1			S36	S8 S18 S55	S19 S81			
Industrial or non-industrial	22		S56	S29 S38	S22 S32	S13 S24 S44	S7 S45 S46	S30 S36	S2 S59	S3 S34 S35	S80	S61 S62	S79
Research methods	20		S25	S11 S26 S38		S13 S14 S44 S41	S31 S39 S58	S36	S59	S3 S73	S63 S74	S86	S76 S84
Otherevidence assessment	9							S28	S17 S18	S51 S81	S65		S64 S67 S77
None	18	S6		S10 S12 S40 S42			S47	S20 S23	S27 S48	S57 S85	S60 S75 S82	S70 S72	S71

Table 8
The quality questions (Dybå and Dingsøyr, 2008) and their application in the SRs (cf. the row 'Quality questions' of Table 7). Q1-11 are the original questions and Q12-15 are extensions.

	S1	S3	S8	S11	S13	S14	S18	S33	S36	S52	S55	S76	S80
Q1 Is the paper based on research (or merely a lessons learned report based on expert opinion)?	x	x	x	x	x					x	x	x	x
Q2 Is there a clear statement of the aims of the research?	x	x	x	x	x	x	x	x	x	x	x	x	x
Q3 Is there an adequate description of the context in which the research was carried out?	x		x	x	x	x	x	x	x	x	x	x	x
Q4 Was the research design appropriate to address the aims of the research?	x		x	x	x		x	x	x	x	x	x	
Q5 Was the recruitment strategy appropriate to address the aims of the research?				x	x					x	x		
Q6 Was there a control group to compare treatments?	x		x	x	x					x			
Q7 Was the data collected in a way that it addressed the research issue?	x	x	x	x	x			x		x		x	
Q8 Was the data analysis sufficiently rigorous?	x	x	x	x	x				x	x	x	x	
Q9 Has the relationship between researcher and participants been adequately considered?	x		x	x	x		x		x	x			
Q10 Is there a clear statement of findings?	x		x	x	x	x	x	x	x	x	x	x	x
Q11 Is the study of value for research or practice?			x	x	x	x				x		x	
Q12 Is there a way to validate the methods?											x		
Q13 Are there any practitioner-based guidelines?	x	x	x								x		
Q14 Are limitation and credibility of the study discusses explicitly							x		x				x
Q15 Was the article refereed?		x											

into one class, but still 19 different research method classes remained. Only S11 (and its extension S13) as well as S36 and S63 use the same categorization, and S73, S74 and S86 rely on the same source (Easterbrook et al., 2008) but use slightly different categorizations. Many research methods are also described briefly or just mentioned by name. Even the primary studies are not always clear about their methods.

Other. Finally, nine SRs use an assessment framework of their own. These frameworks intend to categorize evidence somewhat as the evidence level and research classification above and have even similar categories. However, these categorizations are unique enough that they are not feasible to synthesize or to combine with above frameworks.

3.6. Data on RQ2.3: Evidence-based findings presented in SRs

Evidence-based findings are made in 39 SRs. The 23 SRs scoring 0.5 in QA7 were already described in Section 3.3. Next, we depict the evidence of the seven SRs scoring 1 in QA7. S12 summarizes practitioner experiences on the scalability of variability models found in four primary studies. S13 synthesizes the best evidence about the positive and negative effects that variability modeling causes, e.g., to the productivity and complexity. S14 has a research question about successful industrial experiences of agile SPLs. Experiences identified in the primary studies are described and some topics are listed along with the supporting evidence. S22 classified the primary studies based on their success in applying a feature model and gives a description of each case. S24 included only the primary studies with industrial experience on SPL testing and describes each of them. S45 identifies and summarizes two primary studies with direct industrial evidence on quality attribute variability, and four additional studies that use industrial SPLs as an example. S66 describes the evidence of all primary studies.

3.7. Data on RQ3.1: Variability models

The variability models covered by each of the 39 SRs containing variability models are presented in Tables 9 and 10. In both tables, the 'ID' column identifies the SR; the 'Variability Model' column identifies a variability model as presented by the SR only with minor modifications, such as minor wording changes. Some of the variability models are just mentioned by a name without additional description or very generic name is given that both resulted in including them as *Other*. The 'Support level' column specifies how many of the included primary studies report the variability model, and the corresponding percentage is in the '%' column. Table 10 summarizes the SRs that do not provide evidence from the primary studies for the variability models while Table 9 summarizes the SRs that provide tracing to the evidence—the evidence is elaborated below in RQ3.2. A variability model is not necessarily covered in the all primary studies of an SR and one primary study can cover several models. For example, S19 explicitly analyzes all its 13 primary studies but finds a variability model only in three, S37 has a dedicated section about variability modeling, and S18 takes a sample of only 20% (17) of the primary studies with the highest quality score.

We also grouped the variability models into 14 groups as shown in Table 11 to facilitate more relevant analysis, summary, and synthesis. The abbreviation of a group is used in the 'Group' column of Tables 9 and 10. The groups emerged by first grouping similar variability models. For example, different kinds of feature-based models were all grouped under 'FM'. Similarly, we formed generalized groups, as in the case of the 'Orthogonal' group, to which we assigned different kinds of orthogonal models, such as CVL (Haugen et al., 2013) and COVAMOF (Sinnema et al., 2004), in addition to Orthogonal Variability Model (OVM) as presented by Pohl et al. (2005). We also looked for variability models that appear in several SRs to constitute a group of their own. For example, despite the fact that use cases can be considered a part of UML, 'Use case' formed a group of its own, while other UML-based models were grouped under 'UML'. The 'Other' group consists of the remaining isolated and unique variability models, which did not appear in more than two SRs, and models for which enough details are not given. Finally, we were looking for models, but some SRs included also source code-level techniques, such as aspects (S13), that constitute the 'SRC' (source code) group. The 'Unspecified' group means that it remained unclear whether variability is modeled at all or the SR itself used a similar term. These last two groups are included mainly for the sake of completeness, as the SRs originally reported them. In addition, Table 11 shows a short description, the number of SRs ('SR') in each group, the average support level ('Avg%'), and the number of primary studies in each a group ('#'). The average support level is calculated as an average of the support levels from the SRs that appear in each group.

The variability model groups are divided into two classes in Table 11. First, three groups ('FM', 'DM' and 'Orthogonal') represent dedicated variability models that are independent from existing models and propose constructs specific to variability. The 'FM' group relies on features as the main concepts and focuses primarily on the features that are relevant in the context of variability. The 'DM' group is based on models of decisions that need to be made for resolving variability. The 'Orthogonal' group provides a set of constructs for a dedicated variability model, e.g., by using variability points, but the variability model is not autonomous as it is always linked to the existing, so-called base models of software, such as UML diagrams. However, orthogonal variability models are ignorant about the base model, so any software model can be a base model. In addition, other groups, such as the 'Other' and 'Formal' groups may contain a few unique dedicated variability models. Second, the rest of the groups rely on the existing models of software that are adapted or extended with variability information. For example, variability stereotypes are specified for UML, or variability-specific structured natural language is used for textual requirements or in use cases.

Finally, we mapped the variability models to the topics presented in RQ1. Table 12 shows the topics and the number of SRs that report variability models of the total number of SRs covering each topic.

3.8. Data on RQ3.2: Evidence about variability models

The variability models for which evidence can be extracted are presented in Table 9 as described above with the additional last column summarizing the evidence for each technique. S1, S36 and S81 apply the evidence levels (cf. Section 3.5). The other SRs are mapped to the evidence levels as detailed in Table 13 in order to make the different forms of evidence more commensurable. The resulting evidence for the variability model groups is shown in Table 14.

4. Summary and further analysis of answers to the research questions

This section discusses the answers to the research questions.

4.1. Analysis of RQ1.1: SR publications

The numbers of SRs as well as their authors and institutions show that SR as a methodology is adopted widely. The first SR was

Table 9
Variability models with possibility to trace the evidence.

ID	Group	Variability model	Support level	%	Evidence
S1	Use case	Use case	11 / 49	22%	L1:2 L2:2 L3:1 L4:1 L5:4 L6:1
	Unspecified	Not specified	1 / 49	2%	L1:0 L2:0 L3:0 L4:0 L5:0 L6:1
	FM	Features	7 / 49	14%	L1:1 L2:0 L3:1 L4:1 L5:3 L6:1
	Other	Other uniques	13 / 49	27%	L1:0 L2:3 L3:3 L4:0 L5:5 L6:2
	NL	Natural language	13 / 49	27%	L1:1 L2:3 L3:2 L4:1 L5:6 L6:0
	DM	Decision model	1 / 49	2%	L1:0 L2:0 L3:0 L4:0 L5:1 L6:0
	FM	Feature model	5 / 49	10%	L1:0 L2:2 L3:1 L4:0 L5:2 L6:0
	UML	UML	2 / 49	4%	L1:1 L2:0 L3:1 L4:0 L5:0 L6:0
	OVM	OVM	2 / 49	4%	L1:0 L2:1 L3:0 L4:1 L5:0 L6:0
	Other	Goals	5 / 49	10%	L1:0 L2:3 L3:1 L4:1 L5:0 L6:0
S3	FM	Features	2 / 13	15%	L1:0 L2:1 L3:0 L4:0 L5:1 L6:0
	FM	Feature model	5 / 13	38%	L1:0 L2:1 L3:0 L4:2 L5:2 L6:0
	NL	Structured natural language sentences	2 / 13	15%	L1:0 L2:1 L3:0 L4:0 L5:1 L6:0
	NL	Classification of requirement sentences	3 / 13	23%	L1:0 L2:2 L3:0 L4:1 L5:0 L6:0
S19	SA	Architecture view for decisions (no details)	1 / 13	8%	L1:1 L2:0 L3:0 L4:0 L5:0 L6:0
	OVM	OVM	1 / 13	8%	L1:0 L2:1 L3:0 L4:0 L5:0 L6:0
	FM	Feature model	1 / 13	8%	L1:0 L2:1 L3:0 L4:0 L5:0 L6:0
S22	FM	Successful FM application with adaptations	2 / 16	13%	L1:0 L2:0 L3:0 L4:0 L5:0 L6:2
	FM	Successful FM application without adaptations	4 / 16	25%	L1:0 L2:0 L3:0 L4:0 L5:0 L6:4
	FM	FM could have helped	3 / 16	19%	L1:0 L2:0 L3:0 L4:0 L5:0 L6:3
	FM	Unsuccessful FM application	2 / 16	13%	L1:0 L2:0 L3:0 L4:0 L5:0 L6:2
S30	Formal	Formal	2 / 37	5%	L1:0 L2:0 L3:0 L4:2 L5:0 L6:0
	UML	Activity diagram	3 / 37	8%	L1:0 L2:1 L3:0 L4:1 L5:1 L6:0
	UML	Sequence diagram	4 / 37	11%	L1:0 L2:3 L3:0 L4:1 L5:0 L6:0
	Use case	Use cases	9 / 37	24%	L1:0 L2:5 L3:0 L4:3 L5:1 L6:0
	FM	Feature model	4 / 37	11%	L1:0 L2:4 L3:0 L4:0 L5:0 L6:0
	OVM	OVM	5 / 37	14%	L1:0 L2:2 L3:0 L4:3 L5:0 L6:0
	UML	UML	1 / 37	3%	L1:0 L2:0 L3:0 L4:1 L5:0 L6:0
	FM	Other Feature based	1 / 37	3%	L1:0 L2:0 L3:0 L4:1 L5:0 L6:0
	SRC	Source code based	3 / 37	8%	L1:0 L2:1 L3:0 L4:0 L5:2 L6:0
	SA	Architecture model	1 / 37	3%	L1:1 L2:0 L3:0 L4:0 L5:0 L6:0
	Other	Metamodels	1 / 37	3%	L1:1 L2:0 L3:0 L4:0 L5:0 L6:0
S36	NL	Natural language	11 / 46	24%	L1:2 L2:7 L3:0 L4:2 L5:0 L6:0
	Formal	Formal	19 / 46	41%	L1:6 L2:9 L3:1 L4:2 L5:0 L6:1
	Other	Service models	5 / 46	11%	L1:1 L2:2 L3:0 L4:2 L5:0 L6:0
	SA	Architecture model	13 / 46	28%	L1:2 L2:8 L3:0 L4:1 L5:1 L6:1
	Other	Ontology based techniques	3 / 46	7%	L1:0 L2:1 L3:1 L4:1 L5:0 L6:0
	MDE	Domain-specific language	2 / 46	4%	L1:0 L2:2 L3:0 L4:0 L5:0 L6:0
	UML	UML	4 / 46	9%	L1:1 L2:3 L3:0 L4:0 L5:0 L6:0
	Other	Other uniques	3 / 46	7%	L1:1 L2:1 L3:0 L4:1 L5:0 L6:0
S50	Use case	Use case	9 / 9	100%	L1:0 L2:7 L3:0 L4:1 L5:1 L6:0
S51	FM	Feature model	9 / 24	38%	L1:2 L2:2 L3:0 L4:2 L5:3 L6:0
	FM	Features	3 / 24	13%	L1:1 L2:0 L3:0 L4:2 L5:0 L6:0
	Formal	Formal	6 / 24	25%	L1:2 L2:1 L3:0 L4:3 L5:1 L6:0
	Scenario	Scenario	6 / 24	25%	L1:0 L2:0 L3:0 L4:2 L5:4 L6:0
	SA	Architecture model	1 / 24	4%	L1:0 L2:0 L3:0 L4:1 L5:0 L6:0
S56	UML	UML	2 / 10	20%	L1:0 L2:0 L3:0 L4:1 L5:0 L6:0
	FM	Feature model	3 / 10	30%	L1:0 L2:0 L3:0 L4:1 L5:2 L6:0
	DM	Decision model	2 / 10	20%	L1:0 L2:0 L3:0 L4:0 L5:2 L6:0
	NL	Textual	2 / 10	20%	L1:0 L2:0 L3:0 L4:0 L5:2 L6:0
	FM	Features	1 / 10	10%	L1:0 L2:0 L3:0 L4:0 L5:1 L6:0
	SRC	Source code based	1 / 10	10%	L1:0 L2:0 L3:0 L4:0 L5:1 L6:0
S80	FM	FM	25 / 54	46%	L1:0 L2:25 L3:0 L4:0 L5:0 L6:0
	UML	UML class diagrams	18 / 54	33%	L1:0 L2:18 L3:0 L4:0 L5:0 L6:0
	Use case	Use case	6 / 54	11%	L1:0 L2:6 L3:0 L4:0 L5:0 L6:0
	Other	Activity diagram	6 / 54	11%	L1:0 L2:6 L3:0 L4:0 L5:0 L6:0
	MDE	DSL	15 / 54	28%	L1:0 L2:13 L3:0 L4:2 L5:0 L6:0
	OVM	OVM	6 / 54	11%	L1:0 L2:5 L3:0 L4:1 L5:0 L6:0
	Other	Goal Modeling	3 / 54	6%	L1:0 L2:3 L3:0 L4:0 L5:0 L6:0
	Other	Ontologies	4 / 54	7%	L1:0 L2:3 L3:0 L4:1 L5:0 L6:0

published in 2007. Since 2009, the number of published SRs per year has remained quite stable although there seems to be a minor increase over the years (Fig. 3).

The 86 SRs were scattered over 34 different publication forums, which supports the earlier observation (S13 and S52) that SPL and variability studies are not published in a few specific forums. Journals (40, 47%) are the most common publication type, but conferences (36, 42%) and workshops (10, 12%) also publish a significant amount of SRs. The highest number of papers were published by Information and Software Technology (20 SRs), which even advocates SRs. SPL and variability specific forums have published a total of 14 SRs: SPLC (5 SRs), SPLC's workshops (6), ICSR (2), and VaMoS (1).

4.2. Analysis of RQ1.2: SRs characteristics

In terms of the SR type, we relied on the self-claims of the authors. Most of the SRs categorize themselves as SLRs (53, 62%), and the rest are Maps. However, we also analyzed and compared these SLRs and Maps. The average, median and standard deviation of number of included primary studies is for SLRs 61, 47 and 71, and for Maps 43, 36 and 34, respectively. There are 13 (15%) SLRs

Table 10
Variability models without possibility to trace the evidence.

ID	Group	Variability model	Support level	%
S2	NL	Textual variability information	12 / 62	19%
	FM	Feature mapping to elements	18 / 62	29%
	FM	Feature model	13 / 62	21%
	SRC	Source code reorganization for variability	14 / 62	23%
S4	FM	FODA feature model	4 / 17	25%
	FM	Extended feature model	3 / 17	20%
	FM	Other feature model	5 / 17	30%
	FM	Feature model notation unclear	4 / 17	25%
S5	MDE	MDE	1 / 29	3%
	FM	Feature model	3 / 29	10%
	Other	Context modeling	2 / 29	7%
S10	FM	Feature model	14 / 34	41%
	DM	Decision model	6 / 34	18%
	Other	Other uniques	12 / 34	35%
	Other	Independent of specific modeling	2 / 34	6%
S13	FM	Feature model	33 / 97	34%
	UML	UML	25 / 97	26%
	SA	Architecture model	8 / 97	8%
	NL	Natural language	6 / 97	6%
	SA	Component model	5 / 97	5%
	Formal	Formal	4 / 97	4%
	Other	X-Frames	4 / 97	4%
	MDE	Domain-specific language	3 / 97	3%
	Other	Ontology based techniques	3 / 97	3%
	SRC	Aspects-orientation	2 / 97	2%
	OVM	Orthogonal variability management	2 / 97	2%
	Other	Other uniques	2 / 97	2%
S15	Use case	Use case	6 / 64	9%
	UML	UML	12 / 64	19%
	Formal	Formal	4 / 64	6%
	DM	Decision model	1 / 64	2%
S17	UML	UML	2 / 19	11%
	FM	Feature model	8 / 19	42%
	Other	Other	1 / 19	16%
	Unspecified	Unspecified	2 / 19	11%
	MDE	DSL	2 / 19	11%
	DM	Decision model	2 / 19	11%
	OVM	CVL	1 / 19	5%
	Other	Other uniques	2 / 19	11%
S18	FM	Feature model	4 / 196	2%
	Other	Rules, conditions	15 / 196	8%
	Other	Variant labels, annotations	7 / 196	4%
	Other	Profiles, e.g., tables	4 / 196	2%
	Scenario	Change scenarios	2 / 196	1%
S21	FM	Variability model about context combined with a feature model	1 / 37	3%
	FM	Feature model	8 / 37	22%
	SA	Koala model	2 / 37	5%
	UML	UML	2 / 37	5%
S23	FM	Feature model	127 / 127	100%
S27	FM	Features	4 / 20	20%
	FM	Feature model	3 / 20	15%
	NL	Natural language	1 / 20	5%
	UML	UML	2 / 20	10%
	FM	Features	4 / 20	20%
	OVM	Orthogonal	2 / 20	10%
	Unspecified	None	6 / 20	30%
S34	FM	Feature model	45 / 77	58%
	Other	Other than features	22 / 77	29%
S35	FM	Feature model	41 / 47	87%
	Other	Sets of constraints	4 / 47	9%
	OVM	OVM	2 / 47	4%
S37	FM	Feature model	14 / 81	17%
	FM	Other feature based	4 / 81	5%
	OVM	OVM	2 / 81	2%
	OVM	Covamof	2 / 81	2%
	UML	UML	5 / 81	6%
	MDE	Domain-specific language	6 / 81	7%
S38	FM	Feature model	6 / 39	15%
	Other	Other models	6 / 39	15%
	Formal	Formal	7 / 39	18%
	Other	Goal-oriented model	3 / 39	8%
	Other	Tree	5 / 39	13%
	NL	Textual	5 / 39	13%
	Other	Other mechanism	8 / 39	21%
	Unspecified	Unspecified	12 / 39	31%

(*continued on next page*)

Table 10 (*continued*)

ID	Group	Variability model	Support level	%
S42	MDE	DSL	3 / 7	43%
	FM	Feature model	1 / 7	14%
	Unspecified	Not specified	1 / 7	14%
	SRC	Source code based	2 / 7	29%
S45	FM	Features	7 / 29	24%
	FM	Attributes of features	3 / 29	10%
	SA	Attributes of components	4 / 29	14%
	Other	Other (variation points, constraints, dependencies)	6 / 29	21%
S47	Scenario	Scenario	9 / 57	16%
	SA	ADL	3 / 57	5%
	Scenario	ATAM / Saam	5 / 57	9%
	SA	Structure such as component model	7 / 57	12%
S52	FM	Feature model	43 / 63	68%
	Other	Other than features	17 / 63	27%
S55	UML	UML	2 / 36	6%
	FM	Features	12 / 36	33%
	Formal	Formal	13 / 36	36%
	MDE	MDE	4 / 36	11%
	SRC	Source code based	5 / 36	14%
	OVM	OVM	1 / 36	3%
	Unspecified	Unspecified	1 / 36	3%
S58	FM	Feature model	3 / 30	10%
	Other	Goals and scenarios	1 / 30	3%
S63	FM	Feature model	28 / 39	72%
	OVM	CVL	2 / 39	5%
	OVM	Orthogonal	1 / 39	3%
	SRC	Source code	3 / 39	8%
	Other	Goal model	1 / 39	3%
	Unspecified	Unclear	4 / 39	10%
S64	FM	Feature model	47 / 47	100%
S69	Other	Goal model	2 / 54	4%
	FM	Feature model	36 / 54	67%
	UML	UML	4 / 54	7%
	OVM	OVM	3 / 54	6%
	OVM	CVL	3 / 54	6%
	Other	Other	14 / 54	26%
S72	FM	Feature model	14 / 25	56%
	OVM	Orthogonal Variability Model	1 / 25	4%
	FM	Feature list	4 / 25	16%
	Other	Other	4 / 25	16%
	Other	Product comparison matrix	1 / 25	4%
S76	FM	Feature model	28 / 60	47%
	UML	UML	6 / 60	10%
	DM	Decision model	4 / 60	7%
	OVM	CVL	1 / 60	2%
	OVM	OVM	1 / 60	2%
	Other	Other	14 / 60	23%
	FM	FM	5 / 60	8%
S81	Other	Aspect model	1 / 37	3%
	Other	Actor model	1 / 37	3%
	Other	MVRP	1 / 37	3%
	Other	OWL	1 / 37	3%
	Other	OCL	1 / 37	3%
	FM	Feature model	18 / 37	49%
S82	UML	UML	18 / 35	51%
	Other	Component-Based Architecture	14 / 35	40%
	Other	Module Diagram	8 / 35	23%
	Other	Other	16 / 35	46%
	Unspecified	Not Specified	3 / 35	9%
S83	FM	Feature model	18 / 35	51%
	Other	Goal model	1 / 35	3%
	Other	Other	2 / 35	6%
	Other	Ontologies	2 / 35	6%
	Other	Dependency models	4 / 35	11%
	Other	Graphs	6 / 35	17%

and 5 (5%) Maps with fewer than 20 papers. The largest number of primary studies in a Map is 423 (S68), which is an exception because only two other Maps have more than 100 primary studies (119 in S61 and 165 in S65). The largest number of primary studies in SLRs is 196 (S18) and four SLRs have more than 100 primary studies. The QA scores also do not indicate significant differences (Table 15). Restricting the publication year of SLRs to the period when Maps were published after the guidelines had been published (Petersen et al., 2008), 2011 and onward, does not significantly change the comparison between SLRs and Maps.

Consequently, our analysis does not show significant differences between Maps and SLRs as they are being applied. The difference between Maps and SLRs is not defined in principle by the number of primary studies, but because maps cover a wider area, one could

Table 11
The groups of variability models and their descriptions, the number of SRs in which the group appears ('SRs'), the average support level percentage for the group ('%'), and the total number of primary studies in SRs for the group ('#').

Group	Description	SRs	Avg%	#
Dedicated variability models:				
FM	Feature diagrams or trees, and other feature-based models.	34	44%	721
Orthogonal	Orthogonal variability models, e.g., CVL and OMV.	14	6%	38
DM	Decision models.	6	10%	16
Extensions and adaptations:				
Other	Models that only a few primary studies apply in an SR and do not appear in several other SRs, or the details of the model are not given.	23	18%	263
UML	Generally, any UML diagram other than use case, including adaptations such as stereotypes.	15	16%	112
NL	Natural language, structured natural language.	8	19%	55
SA	Software architecture view-based models and component models other than UML and MDE.	8	12%	45
Formal	Formal approaches. Further details are typically not given.	7	19%	55
MDE	Model-driven engineering including domain specific languages (DSL).	8	14%	36
SRC	Source code-level techniques	7	13%	30
Unspecified	Variability model is not specified or applied.	8	14%	30
Use case	Use case with, e.g., variability specialization or adaptation.	5	33%	41
Scenario	Scenario-based models	3	17%	22

Table 12
Variability models for different topics. The '#'column give the number of SRs reporting a variability model out of the SRs in the group and the 'SRs' column lists the SRs.

Topic	#	SRs
Requirements engineering	4 / 7	S1 S3 S50 S81
Design	3 / 6	S19 S56 S47
Testing	5 / 13	S15 S30 S35 S52 S80
Variability management	11 / 25	S4 S10 S13 S18 S22 S23 S34 S64 S69 S76 S82
Quality attributes	5 / 8	S36 S38 S45 S55 S63
Process model	0 / 4	
Management	6 / 17	S2 S3 S27 S51 S72 S83
Specific SPLs	10 / 15	S5 S17 S21 S36 S37 S42 S52 S58 S69 S82
Research methods	1 / 4	S13

Table 13
Evidence level mappings in SRs for evidence about variability models.

ID	Evidence level	Original evidence
S3	L5	Industrial settings with practitioners.
	L4	Industrial settings with researchers.
	L2	Researchers' evaluation.
S19	L2	Example.
	L1	No evidence.
S22	L6	Includes only industrial practice
S30	L5	Industrial experiment.
	L4	Academic experiment.
	L2	Example.
	L1	No evidence.
S50	L5	Validation research.
	L4	Evaluation research.
	L2	Solution proposal.
S51	L5	Industry evaluation.
	L4	Open source or academic.
	L2	Generated.
	L1	None.
S56	L5	Industry.
	L4	No industry.

Table 14
The total number of primary studies in SRs from L1 weakest to L6 strongest level of evidence for the variability model groups.

Variability model group	L1	L2	L3	L4	L5	L6
FM	4	36	2	9	14	12
DM	0	0	0	0	3	0
Formal	7	10	1	7	1	1
MDE	0	15	0	1	0	0
NL	3	13	2	4	9	0
Other	3	22	5	6	5	2
Orthogonal	0	9	0	5	0	0
SA	4	8	0	2	1	1
Scenario	0	0	0	2	4	0
SRC	0	1	0	0	3	0
UML	2	25	1	4	1	0
Unspecified	0	0	0	0	0	1
Use case	2	20	1	5	6	1

Table 15
QA averages in Maps and SLRs. See Table 6 for full QA scores.

	QA1	QA2	QA3	QA4	QA5	QA6	QA7	sum
SLR no (0)	1	6	12	1	34	18	30	
SLR partially (.5)	12	16	24	2	6	9	17	
SLR yes (1)	40	31	17	50	13	26	6	
Avg	0.87	0.74	0.55	0.96	0.30	0.58	0.27	4.26
Map no (0)	0	3	7	0	31	6	17	
Map partially (.5)	8	10	20	3	1	14	15	
Map yes (1)	25	20	6	30	1	13	1	
Avg	0.88	0.76	0.48	0.95	0.05	0.61	0.26	3.98

expect significantly more primary studies in Maps. One could also expect Maps to have lower QA scores because Maps provide an overview of a topic area without performing as in-depth analysis as SLRs do. However, neither of these expectations was realized. In fact, the classification to SLRs and Maps by the authors of SRs is not necessary consistent. The research questions of a Map could cover a wider topic area than that of an SLR, but we did not analyze this aspect.

In general, the distinction between Maps and SRs would be better realized and more meaningful if the SLRs were more clearly focused on rigorous analysis taking into account quality of and evidence in the primary studies. Maps would then carry out overviews of wide topic areas and could even rely on shallower

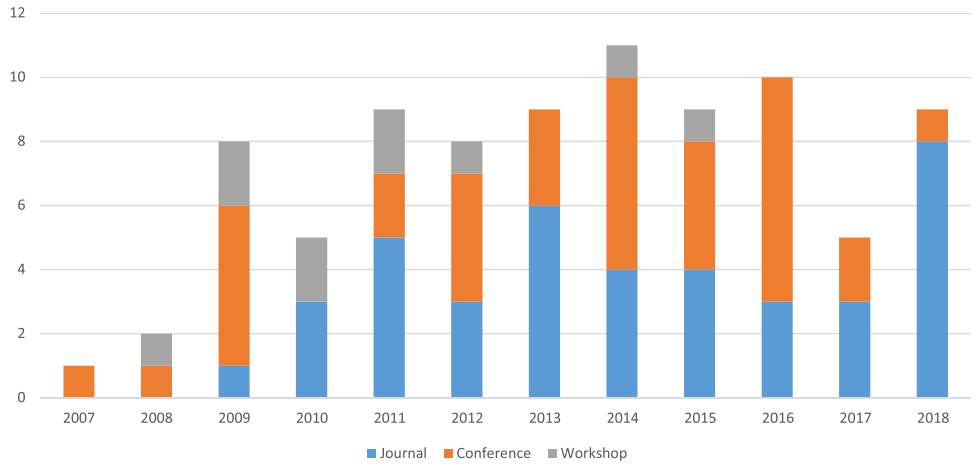

Fig. 3. The distribution of SRs by year and publication type. The search did not cover entire year 2018.

analysis, such as providing only numerical summaries of evidence. These differences are now realized in only some of the SRs.

Due to few and inconsistent differences between SLRs and Maps according to categorization based on what SRs self-claim, the rest of this study does not differentiate between these two.

4.3. Analysis of RQ1.3: Addressed topics

The SRs cover an extensive range of topics around SPLs and variability. The topics are mostly unique and only a few SRs are extended or updated, which are all always done at least partially by the same authors. Some SRs seem to be closely related, such as S15 and S43, although their research questions are formulated differently. However, it is infeasible to study updates, extensions, or trends about the same topic.

The topic categories show that SRs cover other SPL development life-cycle phases except implementation: There is no SR about realizing variability or variants although, e.g., variability management is extensively addressed. Our perception is that detailed design has not been addressed as extensively as high-level architecture design or feature modeling. When considering variability models, our findings are similar to existing studies (Heradio et al., 2016; Marimuthu and Chandrasekaran, 2017) that have found out variability management including feature modeling, and testing to be the dominant topics. However, some of the SRs are each quite specific, focusing on an individual subtopic of emerging areas of research. The SRs of each topic category are generally distributed relatively evenly over the years but the SRs have different focuses within the topics rather than being updates. The most notable exception is the process model category: Both S14 and S53 focus on agile methods and appear in the same journal issue. The latest SR that we classified as research methods was published in 2011. However, many other SRs address research methods—or the poor quality of research method used—in the primary studies. Many SRs of different topics share a common interest in tool support: Four SRs have tools as the main topic, 14 SRs have a research question about tools, and 24 SRs have tools as an analysis topic or field in data extraction.

SRs often disregard, or are implicit about, the central distinction between application and domain engineering. However, for certain management topics, such as adoption strategies (S16), the distinction is not even meaningful but we did not explicitly differentiate SPL management, e.g., similarly as Clements and Northrop (2001). An explicit distinction between application and domain engineering was made in 28 (33%) SRs. To characterize how the SRs address application (AE) and domain (DE) engineering phases, we assigned scores 0, 0.5 and 1 for the extracted data (Table 4) and calculated the sums of scores. In total, domain engineering scored 62.5 and application engineering scored 44, meaning that application engineering has received less attention than domain engineering. Although the difference does not seem to be especially significant, SRs should be more explicit about the distinction that was not always easily evident in our analysis.

As a summary, variability realization is the only topic that seems to be lacking an SR. However, the topics are sometimes covered by quite specific SRs for an emerging research area where mainly solution propositions exist rather than research results with convincing empirical evidence worth analyzing. Thus, several subtopics can still lack an SR that might be worth summarizing as a Map. However, it may also be justifiable to conduct integrating, updating, and synthesizing SRs, especially SLRs, about more general or mature topics, embracing the evidence (cf. RQ2) so that already the inclusion and exclusion criteria make selections in favor of studies with convincing empirical evidence of good quality.

4.4. Analysis of RQ1.4: Quality assessment of SRs

The distribution of scores for the QAs is summarized in Fig. 4. In the data, there are no significant changes or trends in the scores over the years. Because scores for QA4 (presenting results) and QA1 (search) are the best, followed by QA2 (inclusion criteria), the state-of-the-practice seems to perform well in searching and describing the primary studies. As indicated by the worse scores for other QAs, there is a room for improvement in analyzing and taking into account the analyzed quality of and evidence found in the primary studies. The results of the analysis bear similarities with the other tertiary study (Marimuthu and Chandrasekaran, 2017) apart from the fact that we extended the analysis beyond QA4. We elaborate each QA below. **QA1: The inclusion and exclusion criteria.** The inclusion and exclusion criteria (QA1) are usually reported appropriately. The most typical deficiency is that only the search string and research questions are explicated, leaving the criteria somewhat ambiguous. However, the QA1 formulation (cf. Table 3) is relatively loose – the presence of criteria is enough for score 1, but the actual criteria is not analyzed, such as whether the criteria are unambiguous and applicable.

QA2: The completeness of the literature search. Despite fulfilling the QA2 criteria, many searches seem to follow a predefined plan strictly, and the search is terminated once the preplanned activities are carried out, without any further considerations. Assessing the search results is largely missing or unsystematic, and taking advantage of the search results of existing SRs is also quite

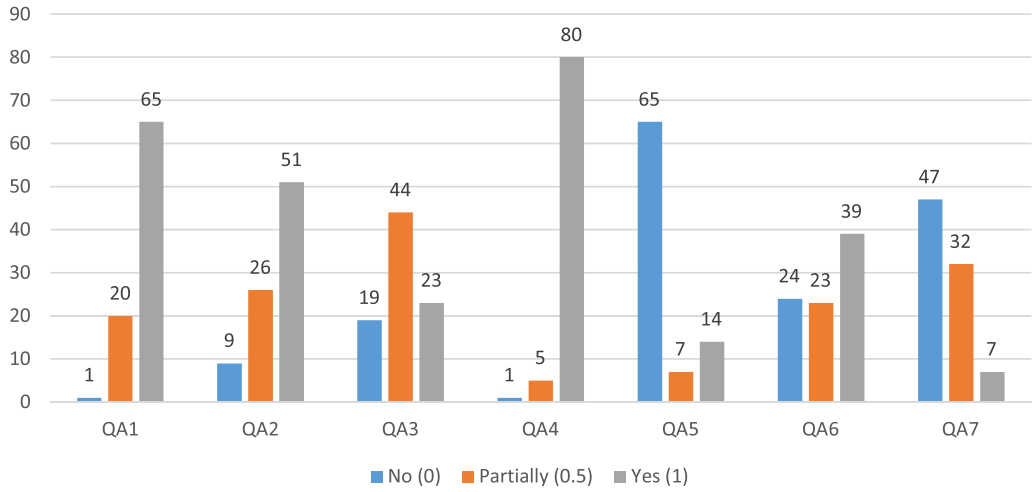

Fig. 4. The distribution of QA scores in the 86 SRs.

rare. For example, if we had been satisfied with only snowballing without the database searches (Fig. 1) in this study, we would have missed a significant number of SRs. However, the QA2 criteria does not require a search validation, such as comparison to other SLRs or using a quasi-gold standard.

The most common shortcoming in the search is relying only on a database search without any augmenting search. Searching a database is unlikely to find anything beyond scientific papers, such as gray literature or books. A snowballing search might be able to find such primary studies better. On one hand, these sources often do not provide scientifically rigorous evidence. On the other hand, if the evidence is not utilized in findings, as indicated by QA7, the restriction is not justified. Limiting the search to a small number of databases or publications is another common shortcoming. For an SPL and for variability (cf. RQ1.2.), in particular, limiting the search to the selected journals and conferences will likely lead to missing primary studies because of scattered publication forums. However, the requirement in QA2 to use four or more databases does not seem to be as clearly justified and would require a more detailed analysis. Further discussion of the search is beyond the scope of this work, and the search for SRs has already been discussed, e.g., in terms of strategies (Kitchenham and Brereton, 2013) and reliability (Dieste et al., 2009; MacDonell et al., 2010; Jalali and Wohlin, 2012; Wohlin et al., 2013).

A recent trend in searches is to carry out snowballing after the introduction of the guidelines (Wohlin, 2014). However, it is not always clear whether snowballing is performed both in backward and forward directions. Another unclear aspect is the extensiveness of snowballing such as whether newly found primary studies are also snowballed.

QA3: The quality assessment of the primary studies. Although several SRs (19) do not take the analyzed quality of or evidence found in the primary studies into account at all, the most common deficiency in many SRs (44) is addressing only the nature of the study and evidence, such as whether industrial evidence is involved or not. These SRs do not assess the primary study quality per se, such as whether the research was designed appropriately and carried out properly. It is also a shortcoming in the QA3 criteria that it combines the quality and evidence, at least as we interpreted it, rather than treats them separately.

QA4: The basic data of the primary studies. The basic studies are typically described adequately, with 80 SRs scoring 1. The space limitations of publications may explain the lower scores. Nevertheless, basic data is required to enable a reader to look at the primary studies for further information.

QA5 & QA6: The quality and evidence traceability of the primary studies. The traceability of the quality (QA5) and evidence (QA6) are dependent on the analyzed quality of and evidence found in primary studies (QA3). The QA5 scores, in particular, are poor due to the lack of a quality analysis. The evidence is analyzed better, resulting in the traceability of evidence (QA6) also being better. The typical limitation of both QA5 and QA6 is that only a numeric summary or a bubble chart is given about the state-of-the-art rather than full traceability.

QA7: The use of quality and evidence in findings. Clear evidence-based findings are scarce, as indicated by the low QA7 scores. The QA7 criteria are relatively loose and require only some kind of elaboration of evidence for the findings. Despite this, only seven SRs score 1. Two typical limitations, which result in a score of 0.5, are that either the primary studies, their topics, and the nature of the evidence are listed, or only a numerical summary of a topic and the related nature of evidence are presented. Neither of these describe actual details of the evidence.

4.5. Analysis of RQ2.1: Quality assessment of primary studies in SRs

The quality questions framework (Dybå and Dingsøyr, 2008) is the most commonly used quality assessment and the only means applied in several (13) SRs (Table 7). Three SRs apply quality as a part of the inclusion criteria, but the criteria either seem general and challenging to measure objectively or not enough detail is provided. The remaining SRs apply their own quality assessment criteria in which the intention does not seem different from the quality questions.

Consequently, the SRs assess the analyzed quality of the primary studies quite rarely, although the quality is important for the validity of evidence-based findings. It remains unclear why so many SRs omit assessment altogether. Despite being the most common framework, even the applicability and value of the quality questions remain unclear because the questions are often adapted or omitted. However, no other means than the quality questions seem to be widely applied. An issue in the application of quality assessment remains to ensure that the assessment is equally applicable to different research approaches and to ensure consistent application. Another issue is whether the value of the quality assessment would be greater when used in the results, or when used as part of the inclusion and exclusion criteria. Nevertheless, SR research would benefit from additional guidelines or examples for quality assessment.

4.6. Analysis of RQ2.2: Evidence assessment of primary studies in SRs

In total, the majority of SRs (62) assess the evidence in the primary studies in at least one way, but there are quite a few different assessments. The assessments that several SRs apply in a similar and commensurable manner are the research classification and evidence level. The differentiation of the study settings and research methods is relatively common, but several different categorizations, which cannot be reliably combined, are used, or the synthesis becomes coarse, such as differentiating industrial studies. In addition, research methods per se do not necessarily say much about the evidence itself.

Although the analysis of evidence seems to be carried out more often and successfully than that of quality, some issues remain. The categorizations or their criteria in an applied framework may be ambiguous or may not be applied uniformly. For example, assessed 'industrial evidence' can be almost insignificant or strong. A highly simplified example that is based on an open source software project or an industrial SPL may bear little actual industrial evidence. Real application in the development of an industrial SPL is much stronger.

S39 highlights an example of interpretation difficulty: A *case study* may refer either to an example or to a rigorous study of the contemporary phenomena in a real context as defined in research methodology literature, such as by Yin (1994). Another issue is that the categorizations are not universally linear. For example, expert opinion (L3) in the evidence-level framework is considered lower than academic studies (L4). Indeed, for some kinds of solutions, such as testing, academic studies, e.g., experiments, can provide better evidence. Some other kinds of solutions, such as SPL adoption, depend on longitudinal activities influenced by a complex context in which better evidence may be obtained even from expert opinion. In addition, expert opinion does not necessarily mean industrial experience. However, the other assessments are even less, if at all, linear. The categorizations focus mostly on empirical evidence, although sometimes, e.g., formal analysis can be a feasible evaluation. We do not argue for complete homogeneity, but a clearer, less fragmented, and more commensurable means of evidence assessment would be beneficial to make SRs more reliable, coherent, and comparable.

4.7. Analysis of RQ2.3: Evidence-based findings presented in SRs.

Evidence-based findings rely on the analyzed quality of and evidence found in primary studies. Only S18, S36 and S88 take the analyzed quality of primary studies into account. These two SRs identify the best primary studies by combining the quality and evidence into a two-dimensional framework to represent rigor and relevance.

For the evidence, which alone is considered more often, one archetypical means is to describe the evidence in primary studies. The description gives some contextual information, although often quite briefly. However, the evidence is not necessarily synthesized at all or at least in a research questions-centric manner, and it remains focused on the primary studies.

Another archetypical use of evidence is that the SRs present only the levels or classes of evidence (cf. RQ2.2). For this, three archetypical representations are used. First, each primary study is listed or tabulated along with its level of evidence. Sometimes, topics and additional information or groupings are provided. The built-in traceability in a list provides a reader with the possibility to study in more detail the evidence of the primary studies, but the details are not readily provided. A synthesis or an overview of the state-of-the-art cannot be made explicit from a list. Second, a chart, such as a bubble plot, represents different topics or research questions and levels of evidence. Compared to a listing, a chart provides a better overview of different subtopics and a summary of the evidence levels. However, the traceability of primary studies is not possible from a chart but needs to be provided elsewhere. Third, the level of evidence is represented as a completely separate concern for an SR, such as in a separate table or a numeric summary, thereby providing the number of studies at each level of evidence. This means that evidence is not directly linked to the findings. Although such a representation can give an overview of evidence about the topic of the entire SR, it cannot give an overview of the specific research questions.

One issue that arises is that, while evidence may show failure, inefficiency, or inapplicability, this is not clear from the SRs that report only the levels or classes of evidence. Although typical evidence in software engineering research involves confirming or demonstrating success or applicability, S22 provides a counterexample in the form of unsuccessful industrial cases. The evidence should also show, e.g., context, and measures of effect in terms of whether there was any effect and how significant the effect was.

The habit of SRs reporting only the level of evidence, not the evidence itself, is somewhat illogical. Consider the following analogy to medicine. An SR would enumerate the kinds of tests that have been conducted for a drug but would not report the drug's effects, whether positive or negative, on the patients' health or any characteristics of patient or the patient's environment. Respectively, in SRs focused on an SPL and variability, it would be important to take into account the analyzed quality of and evidence found in the primary studies rather than just the levels of evidence. However, in a Map that provides an overview of a research topic, the level of evidence may be sufficient. If evidence is not considered at all, a reader cannot distinguish between, e.g., an idea that has not been concretely tested, a solution proposal that has been tested for feasibility, and a solution that is in industrial use and has proven to be effective.

Finally, an increasingly popular trend, which became more evident during the search update, is the use of SRs as a multi-method or triangulation manner. An SR is combined in a study with another research method, such as a case study or survey, and used as one data source to extend, refine or evaluate the study propositions or findings (Bastos et al., 2017; Da Silva et al., 2015; Myllärniemi et al., 2016; Rabiser et al., 2010; Hohl et al., 2018; Tüzün et al., 2015).

4.8. Analysis of RQ3.1: Variability models

We identified tens of unique variability models from the 39 SRs that reported a variability model (cf. Tables 9 and 13). While some of the variability models were clearly different, there are additionally many minor revisions, such as different representations of or extensions for a feature model. In fact, much of the research on variability models seems to be about proposing new or revised variability models. Separating all the minor differences would result in an even larger set of variability models. Due to the large number of models, some of which have quite small differences, we provided a more meaningful grouping of the variability models. The large variety is further evidenced by the fact that the 'Other' group is the second most common group, appearing in 23 SRs (Fig. 5), having the second largest average support level of 18%, and appearing in the 283 primary studies of the SRs. Only the goal-oriented model appears more than once in the 'Other' group, although some variability models are not described in enough detail to be adequately assessed. However, a feature model is clearly the most commonly appearing variability model (Fig. 5). 'FM' group also stands out among different variability models by having 721 supporting primary studies in the SRs, while the number of supporting primary studies in the next largest groups, 'Other' and 'UML', are 263 and 112, respectively. Even among the different

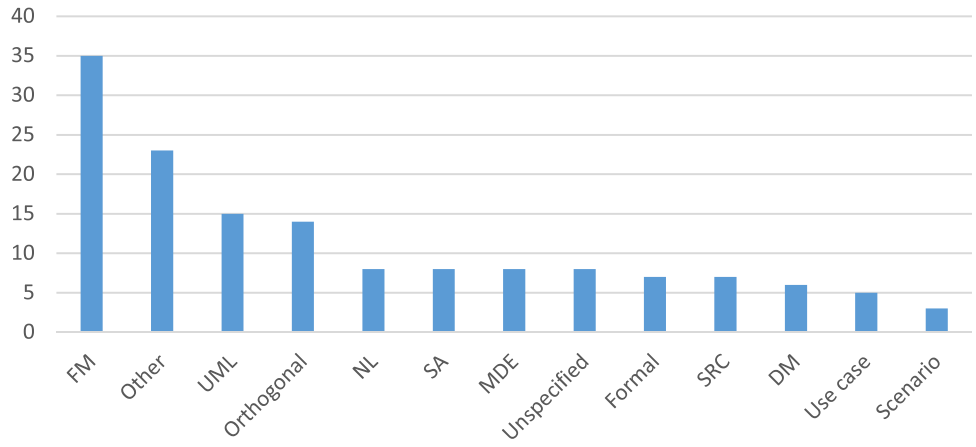

Fig. 5. The number of SRs in which each variability model group appears out of the 30 SRs.

groups, the 'FM' group is mentioned in most (34/39) of the SRs. Out of the five SRs that did not include the 'FM' group, S47 is about architecture evaluation mostly based on existing scenario-based methods, S82 focuses on architecture model, and S50 focuses only on use cases as variability models. The remaining two SRs (S15, S36) did not find anything for the 'FM' group even though the topics are not restricting. Even within SRs, the 'FM' group is dominant, because the support level is, on average, 45%. Four SRs study only a feature model, meaning the support level is 100%, which increases the average support level. However, excluding these four SRs still results in an average support level of 35%.

In addition, UML stands out as another quite often appearing variability model in its own right and as a group, especially if the 'Use case' group is considered a part of UML. Although UML consists of several different diagrams, such as class and sequence diagrams, the 'UML' group is still quite strictly defined compared to any group other than the 'FM' group. In fact, the remaining commonly appearing groups — such as natural language and formal models without further details — are relatively general groups rather than specific models.

Apart from the 'FM' group, the dedicated variability models do not seem to be especially common. Although the 'Orthogonal' group is addressed in 14 SRs, the average support level is low (6%). The 'DM' group is more frequent in early SRs but has been addressed rarely and by only few primary studies in more recent SRs, resulting in an average support level of 10% and 16 primary studies in total.

There are eight specific groups of extensions and adaptations, and most of these groups are addressed by several SRs. Apart from the 'Use case' group, the support level of these groups is, on average, 10–20%, with quite similar standard deviations. Consequently, these groups appear quite frequently, even though their support level does not indicate that they are especially popular.

The extracted variability models cover relatively extensively the different topics of SPL and, thus, the life-cycle phases and activities of SPL development. That is, out of the topics identified in RQ1, all other topics except 'Process model' are covered, and research methods are covered by only one SR. However, a variability model is not especially relevant for either of these two topics. Overall, the support levels of variability models within each topic follow roughly the general averages in Table 11, as discussed above. There are, however, some exceptions to the averages because a topic-specific model is adapted or extended to cover variability, such as use cases for requirements engineering. The group 'FM' is equally dominant for each topic, as it is in general.

Consequently, there seems to be little uncertainty about the feasibility of using some of the existing variability models to express variability as a phenomenon in most of the contexts. Rather, it is a challenge to make a selection and identify differences between the different variability models. The practical needs and use of variability modeling, as well as comparing and synthesizing studies, seem to be more timely issues than any additional novel variability model per se.

Our results bear similarities with the results of SRs that focus on variability. However, S13 identifies significantly fewer models and found UML to be almost as frequent as feature modeling. S18 introduces five abstract categories on the basis of a sample. Only one sample study per category and its central ideas are shown, but the support level is not provided beyond the sample. Consequently, a wider spectrum of variability models is captured by the analysis of this tertiary study, and analysis of variability models with respect to different topics is also provided.

4.9. Analysis of RQ3.2: Evidence about variability models

The identifiable evidence about variability models remains scarce in numbers. Out of the 86 SRs, the variability models are extracted from 39 SRs. However, only nine SRs are reported in a way that provides the variability models with evidence in a traceable manner for a tertiary study. However, as argued above in RQ2, the SR report little evidence that reflects transitively on the found evidence of variability models. Because of the low number of primary studies that provide evidence for variability models, it is more relevant to inspect the evidence in more detail, focusing on practical applications.

The variability models have only isolated primary studies that provide empirical evidence of actual industrial use (L6). The greatest amount of evidence on the actual use of variability models in an industrial context is available for the 'FM' group. S22, which specifically focuses on finding practical applications of feature diagrams, provides 11 out of the 12 primary studies that are classified as L6. However, two of these primary studies are about unsuccessful applications, and three remain hypothetical in the sense that a feature model could have been applied in the situation. The evidence does not seem to be maturing either, because SRs from 2008 and 2010 show the best industrial evidence for 'FM' rather than more recent ones. L6 support is also available for the 'Formal', 'SA', 'Use case' and 'Unspecified' groups, each of which has only one supporting primary study, and the 'Other' group, which has two supporting primary studies. Five of these are reported in S1, and the remaining two are reported in S36. In the end, only three SRs reported evidence on the L6 level.

The L5 evidence level is also based on industrial studies, but these industrial studies typically seem to mean some kind of tests

or informal experiments, such as feasibility tests, performed in an almost realistic environment at best, rather than actual use. Nevertheless, there is more evidence for L5 than for L6, and only the 'MDE' and 'Orthogonal' groups lack evidence on the L5 and L6 levels. Industrial use can also be evidenced at the L3 level by 'Expert opinions' although it is unclear whether the experts are practitioners. Evidence on the L3 level is typically even scarcer than in the case of L6 and mostly originate from S1.

The remaining evidence levels do not include real industrial settings. Therefore, a more detailed analysis between L2 and L4 does not seem reasonable. Besides, the distinction between L2 and L4 does not always seem to be unambiguous or consistent between SRs. Several SRs have primary studies without evidence (L1), and when considering the number of primary studies, only the group 'Formal' has significantly high support level percentages. However, rather than evidence about use, the group 'Formal' can be provided with, e.g., formalisms or proofs that have little merit at the evidence levels.

In sum, the evidence on variability models, especially in practical settings, is numerically scarce and immature. In particular, despite the popularity of the topic and the significant research interest, the evidence on the group 'FM' does not clearly outperform other variability models when the quality and total number of the primary studies are taken into account.

The immature state of the evidence may have different reasons. First, SRs often seem to focus on topics or are carried out in a manner that results in a lot of emphasis being placed on solution proposals rather than specifically focusing on studies about existing industrial practices reported in, e.g., experience reports, surveys, or case studies about variability management. In fact, most of the best evidence is presented in two SRs (S1 and S22).

Second, it may be that few primary studies actually do study variability in practical settings, as indicated by the SRs specifically focusing on industrial experience. Rather, research in primary studies is about variability models as research proposals that do not have evidence of wider industrial adoption, which is then reflected transitively in SRs. Some of the variability models are tried out in industrial settings, but evidence about wider industrial adoption and practice remains scarce. Although research has been conducted, especially for some variability models — to the extent that variability models seem to have matured to a level at which wider industrial adoption could take place — convincing evidence about the transfer of the proposed variability models to wider industrial practice is not common. In particular, the popularity of the 'FM' group indicates that most solution proposals are based on feature models commonly found in academic studies that explore variability as a phenomenon.

Third, because of these focuses, primary studies or SRs, or both, seem not to cover or even consider in-house or informal variability models and other approaches. Only specific variability models are searched for, with no reliable consideration of how often and in which contexts a variability model is considered a necessity and variability is managed by other means. Evidence from observational studies is needed to enable a deeper understanding of the practical application of variability models.

5. Limitations and threats to the validity

To ensure the search coverage, we applied several search strategies. However, one limitation of the search coverage is that we did not contact SRsâ authors or manually search their publication records for additional SRs. Another limitation is that our manual search was not extensive but focused only a limited set of conferences and omitted journals devoted to SRs. However, we initially designed only snowballing search but we already extended the design to cover database searches so that we had reached a saturation point at which the manual searches were no longer discovering any potential new papers for inclusion. Therefore, we do not consider the scope of our manual searches to be a major threat. A snowballing search with Scopus missed some SRs as a result of incorrect or missing records. For the observed issues in Scopus, we applied Google Scholar Web search. Google Scholar resolved the issues, but one major drawback was the large number of false positives.

For the search terms to find SRs from databases, we combined all the terms associated with SRs that we found in the known eleven tertiary studies. However, unlike for snowballing, we omitted the assessment of the database search results. While the search terms about SRs seem relatively comprehensive, one threat involves restricting terms only to SPL and variability, and their synonyms. Phenomena similar to variability exist in domains such as software reuse, autonomous systems, pervasive systems, and service-oriented systems. In addition, potentially relevant search terms for variability, such as personalization, customization, or adaptation, were not included in our search string. The used search terms were also a design decision to limit the context mainly to SPL in which variability, as a research topic, has a central role. There is also a threat that even if a SR was found, the manual application of inclusion and exclusion criteria was applied incorrectly. To alleviate this threat, we carefully analyzed and discussed all borderline SRs.

Data extraction was a manual process that could potentially have resulted in inaccuracies. To mitigate this threat, we applied existing criteria from other tertiary studies, discussed the criteria and borderline interpretations, and re-checked the SRs when in doubt. Some data has dependencies such as QA3 score and applied analysis frameworks. We extracted such data first and then checked the dependencies to validate the data.

A threat with respect to findings is that, as a tertiary study, we relied on SRs and did not analyze the primary studies of the SRs. Any mistake made in the SRs is transitively reflected in our findings. In addition, although the topics of SRs differ, some of the SRs are closely related and can include the same primary studies. Thus, the same primary study may affect our findings several times. To mitigate these threats, we tried to be cautious whenever only minor indications existed. The data extraction was not entirely checked by more than one person, which remains a threat.

The findings and generalizations we made are specific to SPLs and variability. However, software engineering based on SPLs and covering variability is similar to other kinds of software engineering such as single system development. Therefore, our findings could be representative even beyond the current scope of the study, providing generalizable hypotheses worth further investigation.

6. Conclusions

Systematic reviews (SRs) have become over the years an established and relatively widely adopted research methodology. However, SLRs and Maps that have recently been published are in many cases practically indistinguishable. Therefore, we strongly recommend the future SR authors to more clearly distinguish between Maps and SLRs. To this end, further research on SR method application and additional guidelines would help to clarify more clearly the distinction between these two kinds of SRs. For example, SLRs could focus more clearly on a rigorous analysis of particular topics areas, emphasizing the analyzed quality of and evidence found in the primary studies already covered by the inclusion and exclusion criteria. Maps, on the other hand, could carry out wider overviews of emerging topic areas or topics that cut across multiple disciplines.

In terms of the characteristics of the existing SRs, they perform relatively well in systematically carrying out the search, selecting primary studies, and presenting the data about the primary studies according to established quality criteria. However, the SRs should more critically examine the search and analysis results. Furthermore, the SRs rarely take into account the analyzed quality of or evidence found in the primary studies in the synthesis, which should be given more emphasis while conducting SRs. The quality and evidence are also relevant for making evidence-based conclusions. There is also a need for additional SR guidelines or criteria, both in terms of quality and evidence assessment. The guidelines should address issues such as omitting analysis; inconsistent analysis across different SRs; habitually adapted and similar, competing criteria; weak conclusions from analysis results; and an inability to cover various kinds of primary studies.

In terms of the experience of carrying out this tertiary study, similar recommendations apply. It was unexpected that snowballing would not result in more satisfactory results, although some SRs were found only by means of snowballing. Search validation on the basis of a quasi-gold standard was not successful either. Therefore, we recommend, for any SR, to pilot different search strategies and to assess the search results critically. In inclusion, more strict selection criteria could be used, because not all SRs are necessarily reliable or exhaustive enough. Similarly, the division between SLRs and Maps, relying only on authors' claims, turned out to be inconsistent. In general, there is a need for clear criteria to distinguish between the two. For data extraction, we should have included, e.g., research challenges from the SRs, although our perception was that SRs call for better evidence and point out specific future work items that the authors themselves will address.

The topics of the found SRs cover SPL and variability relatively broadly, except for variability realization, which is a potential topic to be covered in a future SR. Because some topic areas are covered by quite specific SRs, there may still be other topics worthy of an SR. It is also worthwhile to update an SR. However, it is more relevant for future SRs to embrace the evidence more — to the extent that only primary studies with realistic evidence are included.

The state-of-the-art of SPL and variability research consists of numerous proposals for a broad spectrum of topics. We did not identify any specific topic that urgently requires novel solutions. Rather, there is a need for actionable solutions that have been evaluated for practical applicability and for suitability and integrability with existing development practices. There is also a need to elaborate and explicate the suitable contexts more clearly, and compare different solutions in different contexts and for different purposes.

As for variability models, we identified a large number of different kinds of variability models, although a feature model stands out clearly as the most popular one and can be considered an archetypical variability model, especially for researchers. Another commonly appearing variability model is UML. Apart from feature models, the dedicated variability models are not be especially widely adopted or gaining in popularity. Several different extensions and adaptations to existing software models exist but are not especially popular. Nevertheless, practically all the phases in software engineering seem to be supported by some kind of variability model, although a feature model is common for any phase. However, the assessed evidence about variability models is, in general, numerically scarce and immature, particularly with regard to industrial use. Consequently, there is little uncertainty about the feasibility of using some of the existing variability models to express variability as a phenomenon in most of the contexts. Rather, one challenge involves making a selection for the context based on understanding the differences and characteristics of the different variability models. The practical needs and use of variability modeling, as well as comparative and synthesizing studies, are more timely research issues than any additional novel variability model.

Acknowledgments

This work was supported by TEKES as part of the Need for Speed (N4S) program; and European Union's Horizon 2020 research and innovation programme under grant agreement no. 732463 as a part of the OpenReq project.

Appendix

Included systematic reviews

S1 V. Alves, N. Niu, C. Alves, and G. Valença, "Requirements engineering for software product lines: A systematic literature review," *Information and Software Technology*, vol. 52, no. 8, pp. 806–820, 2010.

S2 W. K. G. Assunção and S. R. Vergilio, "Feature location for software product line migration: A mapping study," in *International Software Product Line Conference (SPLC) - Volume 2*, 2014, pp. 52–59.

S3 N. Bakar, Z. Kasirun, and N. Salleh, "Feature extraction approaches from natural language requirements for reuse in software product lines: A systematic literature review," *Journal of Systems and Software*, vol. 106, pp. 132–149, 2015.

S4 C. Bezerra, R. Andrade, and J. Monteiro, "Measures for quality evaluation of feature models," in *International Conference on Software Reuse (ICSR)*, 2014, pp. 282–297.

S5 V. A. Burégio, S. R. de Lemos Meira, and E. S. de Almeida, "Characterizing dynamic software product lines-a preliminary mapping study." in *International Software Product Line Conference (SPLC) - Volume 2*, 2010, pp. 53–60.

S6 D. Cabrero, J. Garzas, and M. Piattini, "Understanding product lines through design patterns," in *International Conference on Software and Database Technologies (ICSOFT)*, 2007.

S7 I. do Carmo Machado, J. D. McGregor, and E. Santana de Almeida, "Strategies for testing products in software product lines," *SIGSOFT Software Engineering Notes*, vol. 37, no. 6, pp. 1–8, Nov. 2012.

S8 I. do Carmo Machado, J. D. McGregor, Y. C. Cavalcanti, and E. S. de Almeida, "On strategies for testing software product lines: A systematic literature review," *Information and Software Technology*, vol. 56, no. 10, pp. 1183–1199, 2014.

S9 D. Castelluccia and N. Boffoli, "Service-oriented product lines: A systematic mapping study," *SIGSOFT Software Engineering Notes*, vol. 39, no. 2, pp. 1–6, Mar. 2014.

S10 L. Chen, M. Ali Babar, and A. Nour, "Variability management in software product lines: A systematic review," in *International Software Product Line Conference (SPLC)*, 2009, pp. 81–90.

S11 L. Chen, M. Ali Babar, and C. Cawley, "A status report on the evaluation of variability management approaches," in *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 2009, pp. 118–127.

S12 L. Chen and M. Ali Babar, "A survey of scalability aspects of variability modeling approaches," in *International Software Product Line Conference (SPLC) - Volume 2*, 2009.

S13 L. Chen and M. Ali Babar, "A systematic review of evaluation of variability management approaches in software product lines," *Information and Software Technology*, vol. 53, no. 4, pp. 344–362, 2011.

S14 J. Díaz, J. Pérez, P. P. Alarcón, and J. Garbajosa, "Agile product line engineering - a systematic literature review," *Software - Practice and Experience*, vol. 41, no. 8, pp. 921–941, 2011.

S15 E. Engström and P. Runeson, "Software product line testing - a systematic mapping study," *Information and Software Technology*, vol. 53, no. 1, pp. 2–13, 2011.

S16 J. Ferreira Bastos, P. A. da Mota Silveira Neto, E. S. de Almeida, and S. R. de Lemos Meira, "Adopting software product lines: A systematic mapping study," in *International Conference*

on Evaluation and Assessment in Software Engineering (EASE), 2011, pp. 11–20.

S17 P. Gadelha Queiroz and R. Vaccare Braga, "Development of critical embedded systems using model-driven and product lines techniques: A systematic review," in *Brazilian Symposium on Software Components, Architectures and Reuse (SBCARS)*, 2014, pp. 74–83.

S18 M. Galster, D. Weyns, D. Tofan, B. Michalik, and P. Avgeriou, "Variability in software systems — a systematic literature review," *IEEE Transactions on Software Engineering*, vol. 40, no. 3, pp. 282–306, 2014.

S19 I. Groher and R. Weinreich, "Variability support in architecture knowledge management approaches: A systematic literature review," in *Hawaii International Conference on System Sciences (HICSS)*, 2015, pp. 5393–5402.

S20 R. Heradio, D. Fernandez-Amoros, J. Cerrada, and I. Abad, "A literature review on feature diagram product counting and its usage in software product line economic models," *International Journal of Software Engineering and Knowledge Engineering*, vol. 23, no. 8, pp. 1177–1204, 2013.

S21 G. Holl, P. Grünbacher, and R. Rabiser, "A systematic review and an expert survey on capabilities supporting multi product lines," *Information and Software Technology*, vol. 54, no. 8, pp. 828–852, 2012.

S22 A. Hubaux, A. Classen, M. Mendonça, and P. Heymans, "A preliminary review on the application of feature diagrams in practice," in *International Workshop on Variability Modelling of Software-Intensive Systems (VaMos)*, 2010, pp. 53–59.

S23 A. Hubaux, T. Tun, and P. Heymans, "Separation of concerns in feature diagram languages: A systematic survey," *ACM Computing Surveys*, vol. 45, no. 4, pp. 51:1–51:23, 2013.

S24 M. Johansen, Ø. Haugen, and F. Fleurey, "A survey of empirics of strategies for software product line testing," in *International Conference on Software Testing, Verification, and Validation Workshops (ICSTW)*, 2011, pp. 266–269.

S25 M. Khurum, T. Gorschek, and K. Pettersson, "Systematic review of solutions proposed for product line economics," in *International Software Product Line Conference (SPLC) - Volume 2*, S. Thiel and K. Pohl, Eds., 2008, pp. 277–284.

S26 M. Khurum and T. Gorschek, "A systematic review of domain analysis solutions for product lines," *Journal of Systems and Software*, vol. 82, no. 12, pp. 1982–2003, 2009.

S27 J. Kim, S. Kang, and J. Lee, "A comparison of software product line traceability approaches from end-to-end traceability perspectives," *International Journal of Software Engineering and Knowledge Engineering*, vol. 24, no. 4, pp. 677–714, 2014.

S28 M. A. Laguna and Y. Crespo, "A systematic mapping study on software product line evolution: From legacy system reengineering to product line refactoring," *Science of Computer Programming*, vol. 78, no. 8, pp. 1010–1034, 2013.

S29 B. P. Lamancha, M. P. Usaola, and M. P. Velthius, "Software product line testing: A systematic review," in *International Conference on Software and Database Technologies (ICSOFT)*, 2009, pp. 23–30.

S30 B. Lamancha, M. Polo, and M. Piattini, "Systematic review on software product line testing," in *Communications in Computer and Information Science*, vol. 170, 2013, pp. 58–71.

S31 J. Lee, S. Kang, and D. Lee, "A survey on software product line testing," in *International Software Product Line Conference (SPLC)*, 2012, pp. 31–40.

S32 L. B. Lisboa, V. C. Garcia, D. Lucrédio, E. S. de Almeida, S. R. de Lemos Meira, and R. P. de Mattos Fortes, "A systematic review of domain analysis tools," *Information and Software Technology*, vol. 52, no. 1, pp. 1–13, 2010.

S33 L. Lobato, T. Bittar, P. c. Neto, I. MacHado, E. e. De Almeida, and S. Meira, "Risk management in software product line engineering: A mapping study," *International Journal of Software Engineering and Knowledge Engineering*, vol. 23, no. 4, pp. 523–558, 2013.

S34 R. E. Lopez-Herrejon, L. Linsbauer, and A. Egyed, "A systematic mapping study of search-based software engineering for software product lines," *Information and Software Technology*, vol. 61, no. 0, pp. 33–51, 2015.

S35 R. Lopez-Herrejon, S. Fischer, R. Ramler, and A. Egyed, "A first systematic mapping study on combinatorial interaction testing for software product lines," in *International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 2015, pp. 1–10.

S36 S. Mahdavi-Hezavehi, M. Galster, and P. Avgeriou, "Variability in quality attributes of service-based software systems: A systematic literature review," *Information and Software Technology*, vol. 55, no. 2, pp. 320–343, 2013.

S37 B. Mohabbati, M. Asadi, D. b. Gašević, M. Hatala, and H. Müller, "Combining service-orientation and software product line engineering: A systematic mapping study," *Information and Software Technology*, vol. 55, no. 11, pp. 1845–1859, 2013.

S38 S. Montagud and S. Abrahão, "Gathering current knowledge about quality evaluation in software product lines," in *International Software Product Line Conference (SPLC)*, 2009, pp. 91–100.

S39 S. Montagud, S. Abrahão, and E. Insfran, "A systematic review of quality attributes and measures for software product lines," *Software Quality Journal*, vol. 20, no. 3–4, pp. 425–486, 2012.

S40 M. B. S. de Moraes, E. S. de Almeida, and S. Romero, "A systematic review on software product lines scoping," in *Experimental Software Engineering Latin American Workshop (ESELAW)*, 2009, p. 63.

S41 C. Moraga, M. Moraga, M. Genero, and M. Piattini, "A systematic literature review on software product line quality," in *International Conference on Software and Database Technologies (ICSOFT)*, vol. 2, 2011, pp. 269–272.

S42 Y. Morais and G. E. Thaís Burity, "A systematic review of software product lines applied to mobile middleware," in *International Conference on Information Technology: New Generations (ITNG)*, 2009, pp. 1024–1029.

S43 P. A. da Mota Silveira Neto, I. do Carmo Machado, J. D. McGregor, E. S. de Almeida, and S. R. de Lemos Meira, "A systematic mapping study of software product lines testing," *Information and Software Technology*, vol. 53, no. 5, pp. 407–423, 2011.

S44 E. Murugesupillai, B. Mohabbati, and D. Gašević, "A preliminary mapping study of approaches bridging software product lines and service-oriented architectures," in *International Software Product Line Conference (SPLC) - Volume 2*, 2011, pp. 11:1–11:8.

S45 V. Myllärniemi, M. Raatikainen, and T. Männistö, "A systematically conducted literature review: Quality attribute variability in software product lines," in *International Software Product Line Conference (SPLC)*, 2012, pp. 41–45.

S46 C. R. L. Neto, P. A. M. S. Neto, E. S. de Almeida, and S. R. de Lemos Meira, "Mapping study on software product lines testing tools," in *International Conference on Software Engineering & Knowledge Engineering (SEKE)*, 2012, pp. 628–634.

S47 E. A. Oliveira Junior, I. M. S. Gimenes, and J. C. Maldonado, "Software product line evaluation: Categorization and evolution over the years," in *International Conference on Distributed Multimedia Systems (DMS)*, 2012, pp. 83–88.

S48 J. Pereira, K. Constantino, and E. Figueiredo, "A systematic literature review of software product line management tools," in *International Conference on Software Reuse (ICSR)*, 2014, pp. 73–89.

S49 R. Rabiser, P. Grünbacher, and D. Dhungana, "Requirements for product derivation support: Results from a systematic literature review and an expert survey," *Information and Software Technology*, vol. 52, no. 3, pp. 324–346, 2010.

S50 I. S. Santos, R. M. C. C. Andrade, and P. A. Santos Neto, "How to describe spl variabilities in textual use cases: A systematic map-

ping study," in *Brazilian Symposium on Software Components, Architectures and Reuse (SBCARS)*, 2014, pp. 64–73.

S51 A. R. Santos, R. P. de Oliveira, and E. S. de Almeida, "Strategies for consistency checking on software product lines: A mapping study," in *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 2015, pp. 5:1–5:14.

S52 R. Santos Rocha and M. Fantinato, "The use of software product lines for business process management: A systematic literature review," *Information and Software Technology*, vol. 55, no. 8, pp. 1355–1373, 2013.

S53 I. F. da Silva, P. A. da Mota Silveira Neto, P. O'Leary, E. S. de Almeida, and S. R. de Lemos Meira, "Agile software product lines: A systematic mapping study," *Software - Practice and Experience*, vol. 41, no. 8, pp. 899–920, 2011.

S54 J. da Silva, F. Pereira da Silva, L. do Nascimento, D. Martins, and V. Garcia, "The dynamic aspects of product derivation in DSPl: A systematic literature review," in *International Conference on Information Reuse and Integration (IRI)*, 2013, pp. 466–473.

S55 L. R. Soares, P. Potena, I. d. C. Machado, I. Crnkovic, and E. S. d. Almeida, "Analysis of non-functional properties in software product lines: A systematic review," in *EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA)*, 2014, pp. 328–335.

S56 E. de Souza Filho, R. de Oliveira Cavalcanti, D. Neiva, T. Oliveira, L. Lisboa, E. de Almeida, and S. de Lemos Meira, "Evaluating domain design approaches using systematic review," in *European Conference on Software Architecture (ECSA)*, 2008, pp. 50–65.

S57 E. Tüzün, B. Tekinerdogan, M. Kalender, and S. Bilgen, "Empirical evaluation of a decision support model for adopting software product line engineering," *Information and Software Technology*, vol. 60, pp. 77–101, 2015.

S58 T. Vale, G. B. Figueiredo, E. S. de Almeida, and S. R. de Lemos Meira, "A study on service identification methods for software product lines," in *International Software Product Line Conference (SPLC) - Volume 2*, 2012, pp. 156–163.

S59 G. Vale, E. Figueiredo, R. Abilio, and H. Costa, "Bad smells in software product lines: A systematic review," in *Brazilian Symposium on Software Components, Architectures and Reuse (SBCARS)*, 2014, pp. 84–94.

S60 U. Afzal, T. Mahmood, and Z. Shaikh, "Intelligent software product line configurations: A literature review," *Computer Standards & Interfaces*, vol. 48, pp. 30–48, 2016.

S61 W. K. G. Assunção, R. E. Lopez-Herrejon, L. Linsbauer, S. R. Vergilio, and A. Egyed, "Reengineering legacy applications into software product lines: A systematic mapping," *Empirical Software Engineering*, vol. 22, no. 6, pp. 2972–3016, 2017.

S62 R. Bashroush, M. Garba, R. Rabiser, I. Groher, and G. Botterweck, "CASE Tool Support for Variability Management in Software Product Lines," *ACM Computing Surveys*, vol. 50, no. 1, 2017.

S63 S. Baumgart and J. Fröberg, "Functional Safety in Product Lines - A Systematic Mapping Study," in *Euromicro conferences on software engineering and advanced applications*, 2016, pp. 313–322.

S64 V. Bischoff, K. Farias, L. J. GonÃ§ales, and J. L. V. Barbosa, "Integration of feature models: A systematic mapping study," *Information and Software Technology*, vol. In Press, 2018.

S65 C. Brink, P. Heisig, and F. Wackermann, "Change impact in product lines: A systematic mapping study," *Communications in Computer and Information Science*, vol. 639, pp. 677–694, 2016.

S66 J. Eleuterio, F. Gaia, A. Bondavalli, P. Lollini, G. Rodrigues, and C. Rubira, "On the dependability for dynamic software product lines: A comparative systematic mapping study," in *Euromicro Conference on Software Engineering and Advanced Applications*, 2016, pp. 323–330.

S67 S. El-Sharkawy, N. Yamagishi-Eichler, and K. Schmid, "Metrics for analyzing variability and its implementation in software product lines: A systematic literature review," *Information and Software Technology*, vol. In Press, 2018.

S68 J. Galindo, D. Benavides, P. Trinidad, A.-M. GutiÃ©rrez-FernÃ¦ndez, and A. Ruiz-CortÃ©s, "Automated analysis of feature models: Quo vadis?" *Computing*, pp. 1–47, 2018.

S69 G. Guedes, C. Silva, M. Soares, and J. Castro, "Variability management in dynamic software product lines: A systematic mapping," in *Brazilian Symposium on Components, Architectures and Reuse Software*, 2015, pp. 90–99.

S70 P. Hohl, J. Ghofrani, J. Münch, M. Stupperich, and K. Schneider, "Searching for common ground: Existing literature on automotive agile software product lines," in *International Conference on Software and System Process*, 2017, pp. 70–79.

S71 J. Klünder, P. Hohl, and K. Schneider, "Becoming agile while preserving software product lines: An agile transformation model for large companies," in *International Conference on Software and System Process*, 2018, pp. 1–10.

S72 Y. Li, S. Schulze, and G. Saake, "Reverse engineering variability from natural language documents: A systematic literature review," in *International Systems and Software Product Line Conference*, 2017, pp. 133–142.

S73 C. Lima, M. Cardoso, C. Chavez, and E. Almeida, "Initial evidence for understanding the relationship between product line architecture and software architecture recovery," in *Brazilian Symposium on Components, Architectures and Reuse Software*, 2015, pp. 40–49.

S74 C. Lima and C. Chavez, "A systematic review on metamodels to support product line architecture design," in *Brazilian Symposium on Software Engineering*, 2016, pp. 13–22.

S75 R. E. Lopez-Herrejon, S. Illescas, and A. Egyed, "Visualization for Software Product Lines: A Systematic Mapping Study," in *IEEEE Working conference on software visualization*, 2016, pp. 26–35.

S76 R. E. Lopez-Herrejon, S. Illescas, and A. Egyed, "A systematic mapping study of information visualization for software product line engineering," *Journal of Software: Evolution and Process*, vol. 30, no. 2, 2018.

S77 M. Marques, J. Simmonds, P. Rossel, and M. Bastarrica, "Software product line evolution: A systematic literature review," *Information and Software Technology*, vol. In Press, 2018.

S78 L. Montalvillo and O. DÃaz, "Requirement-driven evolution in software product lines: A systematic mapping study," *Journal of Systems and Software*, vol. 122, pp. 110–143, 2016.

S79 L. Ochoa, O. GonzÃ¦lez-Rojas, A. Juliana, H. Castro, and G. Saake, "A systematic literature review on the semi-automatic configuration of extended product lines," *Journal of Systems and Software*, vol. 144, pp. 511–532, 2018.

S80 M. Sahid, A. Sultan, A. Ghani, and S. Baharom, "Combinatorial interaction testing of software product lines: A mapping study," *Journal of Computer Science*, vol. 12, no. 8, pp. 379–398, 2016.

S81 S. Sepúlveda, A. Cravero, and C. Cachero, "Requirements modeling languages for software product lines: A systematic literature review," *Information and Software Technology*, vol. 69, pp. 16–36, 2016.

S82 L. M. P. da Silva, C. I. M. Bezerra, R. M. C. Andrade, and J. M. S. Monteiro, "Requirements Engineering and Variability Management in DSPLs Domain Engineering: A Systematic Literature Review," in *International Conference on Enterprise Information Systems*, 2016, pp. 544–551.

S83 Z. T. Sinkala, M. Blom, and S. Herold, "A mapping study of software architecture recovery for software product lines," in *European Conference on Software Architecture*, 2018, pp. 49:1–49:7.

S84 L. R. Soares, P.-Y. Schobbens, I. do Carmo Machado, and E. S. de Almeida, "Feature interaction in software product line engineering: A systematic mapping study," *Information and Software Technology*, vol. 98, pp. 44 –58, 2018.

S85 E. Tüzün and B. Tekinerdogan, "Analyzing impact of experience curve on ROI in the software product line adoption process," *Information and Software Technology*, vol. 59, pp. 136–148, 2015.

S86 T. Vale, E. S. de Almeida, V. Alves, U. Kulesza, N. Niu, and R. de Lima, "Software product lines traceability: A systematic mapping study," *Information and Software Technology*, vol. 84, pp. 1–18, 2017.

References

Abran, A., Moore, J.W., Bourque, P., Dupuis, R., Tripp, L.L., 2004. Guide to the software engineering body of knowledge: 2004 version SWEBOK. IEEE Comput. Soc..

Alves, V., Niu, N., Alves, C., Valença, G., 2010. Requirements engineering for software product lines: a systematic literature review. Inf. Softw. Technol. 52 (8), 806–820. doi:10.1016/j.infsof.2010.03.014.

Bano, M., Zowghi, D., Ikram, N., 2014. Systematic reviews in requirements engineering: atertiary study. In: International Workshop on Empirical Requirements Engineering, pp. 9–16. doi:10.1109/EmpiRE.2014.6890110.

Bastos, J., da Mota Silveira Neto, P., O'Leary, P., de Almeida, E., de Lemos Meira, S., 2017. Software product lines adoption in small organizations. J. Syst. Softw. 131, 112–128.

Benavides, D., Segura, S., Ruiz-Cortés, A., 2010. Automated analysis of feature models 20 years later: a literature review. Inf. Syst. 35 (6), 615–636. doi:10.1016/j.is.2010.01.001.

Bosch, J., 2000. Design and Use of Software Architectures: Adapting and Evolving a Product-Line Approach. Addison-Wesley.

Budgen, D., Drummond, S., Brereton, P., Holland, N., 2012. What scope is there for adopting evidence-informed teaching in SE? In: International Conference on Software Engineering, pp. 1205–1214.

Clements, P., Northrop, L.M., 2001. Software Product Lines: Practices and Patterns. Addison-Wesley.

Cruzes, D.S., Dybå, T., 2011. Research synthesis in software engineering: a tertiary study. Inf. Softw. Technol. 53 (5), 440–455. doi:10.1016/j.infsof.2011.01.004. Special Section on Best Papers from XP2010

Da Silva, F.Q., Santos, A.L., Soares, S., França, A.C.C., Monteiro, C.V., Maciel, F.F., 2011. Six years of systematic literature reviews in software engineering: An updated tertiary study. Inf. Softw. Technol. 53 (9), 899–913. doi:10.1016/j.infsof.2011.04.004.

Da Silva, I., Da Mota Silveira Neto, P., O'Leary, P., De Almeida, E., De Lemos Meira, S., 2015. Using a multi-method approach to understand agile software product lines. Inf. Softw. Technol. 57 (1), 527–542. doi:10.1016/j.infsof.2014.06.004.

Dieste, O., Grimán, A., Juristo, N., 2009. Developing search strategies for detecting relevant experiments. Empirical Softw. Eng. 14 (5), 513–539. doi:10.1007/s10664-008-9091-7.

Dybå, T., Dingsøyr, T., 2008. Strength of evidence in systematic reviews in software engineering. In: ESEM, pp. 178–187.

Dybå, T., Kitchenham, B., Jørgensen, M., 2005. Evidence-based software engineering for practitioners. IEEE Softw. 22 (1), 58–65. doi:10.1109/MS.2005.6.

Easterbrook, S., Singer, J., Storey, M.-A., Damian, D., 2008. Selecting Empirical Methods for Software Engineering Research. In: Guide to advanced empirical software engineering. Springer, pp. 285–311.

Galster, M., Weyns, D., Tofan, D., Michalik, B., Avgeriou, P., 2014. Variability in software systems — a systematic literature review. IEEE Trans. Softw. Eng. 40 (3), 282–306. doi:10.1109/TSE.2013.56.

Hanssen, G., Smite, D., Moe, N., 2011. Signs of agile trends in global software engineering research: a tertiary study. In: International Conference on Global Software Engineering Workshop, pp. 17–23. doi:10.1109/ICGSE-W.2011.12.

Haugen, O., Wasowski, A., Czarnecki, K., 2013. Cvl: common variability language. In: Software Product Line Conference. ACM, p. 277. doi:10.1145/2491627.2493899.

Heradio, R., Perez-Morago, H., Fernandez-Amoros, D., Cabrerizo, F.J., Herrera-Viedma, E., 2016. A bibliometric analysis of 20 years of research on software product lines. Inf. Softw. Technol. 72 (Supplement C), 1–15. doi:10.1016/j.infsof.2015.11.004.

Hohl, P., Stupperich, M., Munch, J., Schneider, K., 2018. Combining agile development and software product lines in automotive: challenges and recommendations. In: International Conference on Engineering, Technology and Innovation doi:10.1109/ICE.2018.8436277.

Ivarsson, M., Gorschek, T., 2011. A method for evaluating rigor and industrial relevance of technology evaluations. Empirical Softw. Eng. 16 (3), 365–395,.

Jalali, S., Wohlin, C., 2012. Systematic literature studies: database searches vs. backward snowballing. In: International symposium on Empirical software engineering and measurement, pp. 29–38.

Kang, K., Cohen, S., Hess, J., Novak, W., Peterson, A., 1990. Feature-Oriented Domain Analysis (FODA) Feasibility Study. Technical Report, CMU/SEI-90-TR-21, ADA 235785. Software Engineering Institute.

Kitchenham, B., 2004. Procedures for Performing Systematic Reviews. Technical Report. Keele University Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1.

Kitchenham, B., 2007. Guidelines for performing systematic literature reviews in software engineering. Technical Report, EBSE-2007-01 version 2.3. Keele University.

Kitchenham, B., Brereton, O.P., Budgen, D., Turner, M., Bailey, J., Linkman, S., 2009. Systematic literature reviews in software engineering a systematic literature review. Inf. Softw. Technol. 51 (1), 7–15. doi:10.1016/j.infsof.2008.09.009.

Kitchenham, B., Brereton, P., 2013. A systematic review of systematic review process research in software engineering. Inf. Softw. Technol. 55 (12), 2049–2075. doi:10.1016/j.infsof.2013.07.010.

Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O.P., Turner, M., Niazi, M., Linkman, S., 2010. Systematic literature reviews in software engineering – a tertiary study. Inf. Softw. Technol. 52 (8), 792–805. doi:10.1016/j.infsof.2010.03.006.

MacDonell, S., Shepperd, M., Kitchenham, B., Mendes, E., 2010. How reliable are systematic reviews in empirical software engineering? IEEE Trans. Softw. Eng. 36 (5), 676–687. doi:10.1109/TSE.2010.28.

Marimuthu, C., Chandrasekaran, K., 2017. Systematic studies in software product lines: atertiary study. In: 21st International Systems and Software Product Line Conference. ACM, pp. 143–152.

Marques, A., Rodrigues, R., Conte, T., 2012. Systematic literature reviews in distributed software development: a tertiary study. In: International Conference on Global Software Engineering, pp. 134–143. doi:10.1109/ICGSE.2012.29.

Myllärniemi, V., Savolainen, J., Raatikainen, M., Männistö, T., 2016. Performance variability in software product lines: proposing theories from a case study. Empirical Softw. Eng. 21 (4), 1623–1669.

Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M., 2008. Systematic mapping studies in software engineering. In: International Conference on Evaluation and Assessment in Software Engineering, pp. 68–77.

Pohl, K., Böckle, G., van der Linden, F., 2005. Software Product Line Engineering: Foundations, Principles, and Techniques. Springer.

Rabiser, R., Grünbacher, P., Dhungana, D., 2010. Requirements for product derivation support: results from a systematic literature review and an expert survey. Inf. Softw. Technol. 52 (3), 324–346. doi:10.1016/j.infsof.2009.11.001.

dos Santos, A., de Farias Junior, I., de Moura, H., Marczak, S., 2012. A systematic tertiary study of communication in distributed software development projects. In: International Conference on Global Software Engineering, p. 182. doi:10.1109/ICGSE.2012.42.

Santos, R., De Magalhães, C., da Silva, F.Q.B., 2014. The use of systematic reviews in evidence based software engineering: a systematic mapping study. In: International Symposium on Empirical Software Engineering and Measurement, p. 53.

Santos, R., da Silva, F.Q.B., 2013. Motivation to perform systematic reviews and their impact on software engineering practice. In: International Symposium on Empirical Software Engineering and Measurement, pp. 292–295. doi:10.1109/ESEM.2013.36.

da Silva, F.Q.B., Santos, A.L.M., Soares, S.C.B., França, A.C.C., Monteiro, C.V.F., 2010. A critical appraisal of systematic reviews in software engineering from the perspective of the research questions asked in the reviews. In: International Symposium on Empirical Software Engineering and Measurement, pp. 33:1–33:4. doi:10.1145/1852786.1852830.

Sinnema, M., Deelstra, S., Nijhuis, J., Bosch, J., 2004. Covamof: a framework for modeling variability in software product families. In: Proceedings of Software Product Line Conference (SPLC), pp. 197–213.

Svahnberg, M., van Gurp, J., Bosch, J., 2005. A taxonomy of variability realization techniques. Softw. Pract. Experience 35 (8), 705–754. doi:10.1002/spe.v35:8.

Tüzün, E., Tekinerdogan, B., Kalender, M., Bilgen, S., 2015. Empirical evaluation of a decision support model for adopting software product line engineering. Inf. Softw. Technol. 60, 77–101. doi:10.1016/j.infsof.2014.12.007.

Verner, J., Brereton, O., Kitchenham, B., Turner, M., Niazi, M., 2014. Risks and risk mitigation in global software development: a tertiary study. Inf. Softw. Technol. 56 (1), 54–78. doi:10.1016/j.infsof.2013.06.005.

Wieringa, R., Maiden, N., Mead, N., Rolland, C., 2006. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. Requirements Eng. 11 (1), 102–107. doi:10.1007/s00766-005-0021-6.

Wohlin, C., 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: International Conference on Evaluation and Assessment in Software Engineering, pp. 38:1–38:10. doi:10.1145/2601248.2601268.

Wohlin, C., Runeson, P., da Mota Silveira Neto, P.A., EngstrÃ¶m, E., do Carmo Machado, I., de Almeida, E.S., 2013. On the reliability of mapping studies in software engineering. J. Syst. Softw. 86 (10), 2594–2610. doi:10.1016/j.jss.2013.04.076.

Yin, R.K., 1994. Case Study Research, 2nd Sage: Thousand Oaks.

Zhang, H., Ali Babar, M., 2013. Systematic reviews in software engineering: an empirical investigation. Inf. Softw. Technol. 55 (7), 1341–1354. doi:10.1016/j.infsof.2012.09.008.

Zhang, H., Ali Babar, M., Tell, P., 2011. Identifying relevant studies in software engineering. Inf. Softw. Technol. 53 (6), 625–637. doi:10.1016/j.infsof.2010.12.010.

Zhou, Y., Zhang, H., Huang, X., Yang, S., Ali Babar, M., Tang, H., 2015. Quality assessment of systematic reviews in software engineering: a tertiary study. In: International Conference on Evaluation and Assessment in Software Engineering, pp. 14:1–14:14. doi:10.1145/2745802.2745815.

Mikko Raatikainen is a researcher at Product Requirements and Architecture Research Group of Aalto University and Empirical Software Engineering Research Group of University of Helsinki. His research interests include software product lines, variability, software architecture, and requirements engineering. He is especially interested in conducting empirical research in industrial settings in which software-intensive systems or services are developed.

Juha Tiihonen, D.Sc.(Tech.) is a senior researcher at the Empirical Software Engineering Research Group of University of Helsinki, Finland. His research interests include software product variability and its management, product and service configuration including modeling, configurators and business processes; requirements engineering, recommendation technologies, and mass customization. His strengths include conceptual modeling and tool support for configuration of physical products and services. Juha serves in program committees of numerous journals, conferences and workshops, organizes workshops and serves a co-editor of special issues. Juha is a co-founder of Variantum oy, a company specializing in product life-cycle management of configurable offerings.

Tomi Männistö is professor of Computer Science at the University of Helsinki and the head the Empirical Software Engineering Research Group. His research interests include product requirements and software architectures, software products, software intensive services, product variability and conceptual modelling. Currently, he is also working on the software engineering paradigm called continuous experimentation in which data from real usage of software is collected and analyzed for decision-making in software development. Tomi Männistö is a member and ex-chair of IFIP Working Group 2.10 Software Architecture and representative for the University of Helsinki in ISERN (International Software Engineering Research Network).